POOL

NAVAL RESEARCH LOGISTICS QUARTERLY

DEPOSITORY
MAY 15 1957

OFFICE OF NAVAL RESEARCH

CONTENTS

	Page
ARTICLES	
Measures of Industry Capacity *Edward Zabel*	229
A Bound on the Use of Inefficient Indivisible Units *M. J. Beckmann and J. Laderman*	245
Variants of the Hungarian Method for Assignment Problems *H. W. Kuhn*	253
A Formulation of the Decision Problem for a Class of Systems *J. K. Hale and Ronald L. Reed*	259
On the Theory of Replacement of Machinery with a Random Failure Time *George H. Weiss*	279
An Inventory Control Bibliography *M. Gourary, R. Lewis, and F. Neeland*	295
NEWS AND MEMORANDA	305
RECENT PUBLICATIONS	307
AUTHOR INDEX FOR VOLUME III	313

VOL. 3, NO. 4 DECEMBER, 1956

NAVEXOS P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

EDITORS

Rear Admiral H. E. Eccles, USN (Retired)
The George Washington University

O. Morgenstern
Princeton University

Commander M. I. Rosenberg, USN
Office of Naval Research

F. D. Rigby
Office of Naval Research

M. E. Rose, *Managing Editor*
Office of Naval Research
Washington 25, D. C.

ASSOCIATE EDITORS

J. S. Campbell, *Office of Naval Research*
E. W. Cannon, *National Bureau of Standards*
T. B. Evans, *Colonel, USA*
P. L. Folsom, *Captain, USN*
M. A. Geisler, *The Rand Corporation*
J. D. Hayes, *Rear Admiral, USN (Retired)*
E. B. Henry, Jr., *Commander, USN*
A. J. Hoffman, *Office of Naval Research, London*
S. Karlin, *Stanford University*
J. Laderman, *Office of Naval Research Br. Office*
A. T. Magnell, *Captain, SC, USN*
W. H. Marlow, *The George Washington University*
J. Marschak, *Cowles Foundation*
T. C. Roberts, *Captain, SC, USN*
H. P. Sachaklian, *Colonel, USAF*
E. K. Scofield, *Commander, SC, USN*
J. S. Skoczylas, *Lt. Colonel, USMC*
E. D. Stanley, *Captain, SC, USN*
C. Stein, Jr., *Captain, SC, USN*
R. M. Thrall, *University of Michigan*
C. B. Tompkins, *University of California*
A. W. Tucker, *Princeton University*
R. E. Williams, *Commander, SC, USN*
M. Wood, *Office of the Asst. Sec. of Defense*
M. A. Woodbury, *New York University*

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Manuscripts and other items for publication should be sent to the Managing Editor at the above address. Authors will receive 50 free reprints of their published papers.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U. S. Government Printing Office, Washington 25, D. C. Subscription Price: $1.50 a year in the U. S. and Canada, $2.00 elsewhere; 50¢ for a single copy. Reprints of an individual article can be purchased in quantities of 100 or more. Requests for the purchase price of reprints of a particular article should be sent to the Superintendent of Documents.

The views and opinions expressed in the articles in this Quarterly are those of the authors and not necessarily those of the Office of Naval Research.

The printing of this publication was approved by the Director of the Bureau of the Budget, July 26, 1955

MEASURES OF INDUSTRY CAPACITY[1]

Edward Zabel
Princeton University

A survey of statistical estimates of industry capacity is made, and some possible uses for the measures are discussed.

INTRODUCTION

In a previous paper the author investigated in rather general terms the meanings and implications of a capacity.[2] In the present study an attempt is made to explore particularly the concepts which underlie statistical estimates of industry capacity.

For a sizable number of American industries measures of capacity are published in trade journals and in various federal government periodicals. In general, for industries whose operations are mechanized and whose activity is governed by durable production equipment, the journals tend to publish estimates of capacity; whereas for industries in an agricultural or services classification, such estimates are rarely made. No comprehensive periodic survey is made of industry capacity measures, so it is not possible, without examining hundreds of trade journals, to determine the number of industries which do publish figures. But some idea of magnitude may be surmised from published studies on capacity measures. In 1953 the Inter-Industry Analysis Branch of the Bureau of Mines compiled capacity estimates for 23 Emergency-Model (192 Industry Classification) industries. The 23 E-M industries are primarily in the minerals field and correspond to 92 SIC (Standard Industry Classification) industries. Around one third of these estimates were derived from published data on capacity, while the others were determined from adjusted output figures.[3] Also in 1953 a study was prepared from published capacity data on 15 industries producing basic materials such as steel ingots, petroleum, and paper.[4] A third study, prepared in the Office of Price Administration in 1941, was concerned with the capacities of 71 manufacturing industries. Approximately two fifths of the estimates were taken from trade journal sources, while the remainder

[1]The research for this paper was carried out under office of Naval Research Contract N6onr-27009. The author is indebted to Professor Oskar Morgenstern of Princeton University for his guidance and suggestions.

[2]Edward Zabel, "On the Meaning and Use of a Capacity Concept," Naval Research Logistics Quarterly, December 1955, pp. 237-249.

[3]Henry Goldstein, Pierre Crosson, and Sidney Sonenblum, Capacity Estimates for 23 E-M Industries, Inter-Industry Analysis Branch, Item No. 36, Bureau of Mines, 1953.

[4]Bert G. Hickman, "Cyclical Variations in Capacity and Capacity Utilization in Selected Industries," Conference on Research in Income and Wealth, October 9-11, 1953.

*Manuscript received June 14, 1956.

were derived from informal estimates of trade association officials and federal government industry experts.[5]

An attempt at comprehensively reporting industry capacities was made as a special inquiry for the United States census of manufactures for 1921 and 1923, but after the latter census the inquiry was dropped. As late as the Census of Manufactures, 1947, only scanty data on industry capacity are to be found. The data on productive capacity presented in the census for 1923 were derived primarily from the following entry on the questionnaire submitted to each manufacturer:

> "What is your estimate of the percentage of your output compared with your possible output if you had such demand as to require full running time?"

From the data obtained in response to this question and presumably from other data, the census authorities compiled a ratio of actual output to maximum possible output for all but one of the 333 industries covered in 1923.[6]

The census authorities were not satisfied with the results of the inquiry, and it was this dissatisfaction which led to the abandonment of similar inquiries in subsequent years.[7] The Census Bureau, however, must be held at least partially responsible for the failure of the survey since, clearly, the question they asked was not completely unambiguous. As the replies indicated that such ambiguity was present, the rational procedure would seem to have been to revise the questionnaire rather than to drop the inquiry entirely.

THE CHARACTERISTICS OF INDUSTRY CAPACITY MEASURES[8]

The published capacity measures, generally, have important common characteristics, but the differences which prevail among them in terms of accuracy of data and the precise concept which is the basis for measurement make it mandatory for the user to be well aware of the peculiarities of each estimate. In common, most of the measures purport to show the maximum amount of output which can be produced during some period of time in the industry with the existing durable production equipment. Since the maximum output is not only a function of the durable equipment, however, an answer to this question cannot be derived without specifying what conditions are assumed to prevail in terms of the complementary inputs, the number of shifts worked by the labor force, and the time, if any, allowed for the maintenance and repair of machinery. These assumptions vary widely from industry to industry and, consequently, the raw measures cannot be considered comparable without adjustments for different sets of

---

[5] Robert A. Solo, Industrial Capacity in the United States, Office for Emergency Management and Office for Price Administration and Civilian Supply, June 1941.

[6] Census of Manufactures, 1923, pp. 1201-1207.

[7] The misgivings were indicated in the following letter from the Chief Statistician for Manufactures, U. S. Bureau of the Census, to the Editor of Food Industries:

"Many manufacturers failed to answer the question at all, and the replies made by others indicated that they had failed to understand the import of the question. The purpose was to ascertain the per cent which the actual output formed of the maximum possible output which could have been produced with the equipment in place; but apparently many manufacturers interpreted the question to refer to the maximum possible output which could have been produced with the equipment in place, together with such additional equipment as could have been installed in the space available. It became necessary, therefore, to do a great deal of estimating in order to assemble the statistics; and, since it is the function of the Bureau of the Census to compile and publish reliable and trustworthy statistics, not estimates of doubtful authenticity, the inquiry was dropped."

Quoted in Edwin G. Nourse and Associates, America's Capacity to Produce, Washington, D. C.: The Brookings Institution, 1934, p. 169.

[8] The findings in this section are based on the author's survey of approximately 35 industry measures of capacity published in trade journals and federal government publications.

assumptions. For the bituminous coal industry, for example, the capacity figure is derived by dividing the actual annual output by the average number of days worked and multiplying this figure by 308 days. The capacity estimate is thus influenced by the operating conditions and the economic factors affecting the supply of labor and other inputs in each year, but it does not allow for strikes, breakdowns, or repairs. The data, moreover, are for operating mines only, though idle mines may easily be opened.[9] A more usual practice is to assume an unlimited supply of labor and raw materials, but with varying conditions as to the time use of the machinery, i.e., adjustments may or may not be made for repairs and maintenance. An example is the wheat flour industry where capacity is computed in the following manner. A daily capacity for machinery is obtained from engineering data on how many barrels of wheat flour can be processed over a twenty-four hour period without regard for labor and other input costs. These figures are reported monthly, and an average of the monthly figures is derived to determine average daily capacity. This average is then multiplied by the number of working days in the year (306 during 1923-1948, and 255 from 1949 forward) to obtain annual capacity. The mills covered in the data produce from 90 to 95 percent of the flour.[10]

A serious drawback of capacity statistics is that frequently the complete set of assumptions which are the basis for the estimates are not explicitly described. Conditions as to the number of shifts in operation, number of hours per shift, working days per year, and shutdown time for repairs and maintenance are often not listed. Another drawback for some purposes is that the published data generally tell us nothing about the age distribution and the comparative efficiency of the production units included in the capacity totals. In the same vein, the data frequently do not tell us if output attributable to idle or obsolescent plants and machinery are included in the estimates.

So far we have discussed assumptions which are necessary to derive capacity estimates for single-commodity, single-process industries. Where the industry produces multiple products or where multiple production processes exist, new problems arise. For the multiple-commodity industry where no assumptions are made concerning the product mix, capacity estimates must be given in terms of a transformation function showing the various maximum combinations of output which can be produced or conceivably in terms of the maximum amount of some input item which can be handled by the fixed equipment. The latter would be an acceptable procedure for most uses only if the maximum input flow does not vary greatly with different product mixes. In petroleum refining, for example, the capacity figure for the distillation process is given in terms of the number of barrels of crude oil run to stills. This appears to be a reasonable method, since in distillation the product mix is very narrowly determined by the characteristics of the crude oil.[11] The method used in the paper industry is to base the capacity estimate on what is called normal expectations as to the grade and weight of paper desired by the market.[12] For multiple-process industries various methods are used. An example is the cement industry where production essentially involves two processes: (a) the production of clinker by baking kilns and (b) the conversion of this clinker into finished cement by grinding mills. Capacities for each process can be assigned to any one

---

[9] Minerals Yearbook, 1949, p. 282. However, in Bituminous Coal and Lignite in 1950, Mineral Market Report No. 2032, Department of Commerce, p. 7, the definition of capacity is based on operations for 280 days. This latter definition was supplied by the American Institute of Mining and Metallurgy.

[10] Facts for Industry Series, M-16A, U. S. Bureau of the Census.

[11] Minerals Yearbook, various issues.

[12] American Paper and Pulp Association, Statistics of the Paper Industry, 1951, p. 53.

plant, and the ratio of these different capacities may vary within wide limits from plant to plant. The cement industry, however, publishes official capacity figures based on only its annual capacity to produce clinker.[13]

Industry measures are generally designated as technological measures of capacity because of their noncost nature, but it is customary to classify these concepts even more finely. Three types of technological measure are frequently referred to: rated capacity, theoretical capacity, and practical capacity. There are no standard definitions for these terms, but in general the distinctions are the following. Rated capacity refers to an engineering rating (or an estimate based on past behavior) of the output potential of machinery during some standard period of time, perhaps an hour, a day, a month, or a year. It assumes an unlimited and uninterrupted flow of labor and raw materials, and, if it is computed for a short period of time, it does not usually adjust for delays due to maintenance and repair. The rated capacity of a plant is generally the basic datum from which a theoretical or practical capacity is calculated. Theoretical capacity refers to the maximum output which can be produced during some period of time, usually a year, with the existing equipment, assuming uninterrupted operation. For example, if the rated capacity is the amount of output which can be produced during a twenty-four hour period, the annual theoretical capacity for the plant would be this figure times the number of working days in the year. While the theoretical capacity thus makes no adjustments for interruptions, the practical capacity adjusts the former estimate for such factors as the maintenance and repair time and the usual number of shifts worked in the industry. The differences in meaning noted between theoretical and practical capacity hold fairly consistently among industries, while rated capacity may coincide with either of the former measures. For example, when rated capacity is computed on the basis of a year it usually coincides with theoretical capacity, but occasionally it is adjusted for necessary interruptions so that it is equivalent to a practical capacity. When using capacity data, one must be alert for such nuances but, as we have indicated, it is not always possible to discover the basis for the estimates.

For users of capacity data another caution should be noted relating to the accuracy of the data. While it is difficult to make generalizations, the author encourages an awareness that capacity estimates are not the type of data which generally have the sanction of the major statistical reporting agencies, such as the Bureau of the Census and the Bureau of Labor Statistics (note the comment in footnote 7). The data usually appear originally in trade journals, and it is difficult to ascertain what methods were used to gather the data and what importance they assume in particular industries. Confidence in the statistics must depend on a knowledge of the reporting procedure and what, if any, strategical value the statistics have in each industry.[14]

---

[13] Minerals Yearbook, various issues.

[14] In reporting on the statistics in 1934, Edwin G. Nourse made the following statement which is probably equally as valid today:

"Unfortunately many industries keep no records of capacity. Even for those that do, the figures that pass current in the trade require the closest scrutiny. This is particularly true of statements of capacity in trade directories or similar commercial publications where the reporting company knows that its statement will be published. Under these conditions the return may be influenced by the fact that it will be read by competitors, by tax assessors, and, not least important, by prospective customers. Sometimes the company is tempted to understate, more often perhaps to overstate, its actual performance. In the case of capacity records collected as an incident to the allocation of markets or the establishment of production quotas, the temptation to make the largest showing possible is strong; and for this reason the capacities reported by the international cartels and by certain American trade groups must be scrutinized with care.

(Continued)

In studies on capacity, a question occasionally asked is whether one can add the capacities of the various plants in an industry to get a meaningful industry measure or whether one can add industry measures to get meaningful figures.[15] We have already implicitly answered this question in describing the concepts which are the bases for capacity measures. In a word, the answer depends on the use to which the data are being put. If the objective is to determine the maximum amount of output which can be produced in an industry under specified conditions as to costs and price, then, clearly, adding technical measures of plant capacity, which depend on an unlimited supply of labor and raw materials, will not provide a meaningful answer. In the same sense, if the objective is to determine the capacity output for the economy, adding together technological measures of industry capacities will not give a meaningful answer. On the other hand, if the problem is to determine an indicator of the size of an industry or the growth of an industry over time, the objections to adding technical plant capacities lose most of their force. In the same sense, if one wished to compare differential rates of growth for different groups of industry, comparing summed groups of industry capacity measures may be useful as long as the nature of the capacity measures is kept clearly in mind. The uses and limitations of capacity measures are discussed in more detail in a later section.

In order to make our discussion on industry measures of capacity more pointed, in the next section we shall describe the estimating procedures used in some important industries.

ESTIMATING PROCEDURES USED IN SOME BASIC INDUSTRIES

1. The Steel Industry

Statistics on the steel industry are originally compiled, in the main, by the American Iron and Steel Institute and published initially in various of the Institute's periodicals, such as Steel Facts and the Annual Statistical Report. Data on the capacity of the steel industry are published annually and are obtained by the Institute from questionnaires sent every year to each firm in the industry. A firm is requested to make individual reports for every plant that it owns, and the coverage includes all plants which are primarily engaged in producing pig iron and ingots for casting.

The total annual steel-making capacity in the United States is announced once a year by the Institute, generally as of January 1st, and sometimes at mid-year as well. This practice was begun at the start of 1914, on the basis of a survey made in late 1913. Prior to this time the only official capacity data were published in 1898, 1901, 1904, and 1907 in the directories of the American Iron and Steel Association, the predecessor of the Iron and Steel Institute. Since the 1914 date capacity figures have been reported annually for steel ingot production, and since 1926 for pig iron production, though capacity estimates for the rolling mills are not presented. The reason given for the latter omission is that estimates cannot be made for rolling

---

[14](Continued)
Again, the presence of surplus capacity may be stressed by those already in a business in order to deter newcomers from breaking into it.

"Even where the returns are treated as confidential by the collecting agency, many chances of error arise. Trade statisticians who have developed excellent records of supply and demand in their respective industries advise us that they have found the measurement of capacity impossible to reduce to a generally satisfactory basis. This is partly because interest in the subject is relatively new and partly because of difficulties inherent in the problem. It is not all uncommon to find that different officers of the same company submit differing estimates of the capacity of a given plant."

Edwin G. Nourse, op. cit., pp. 23-24.

[15]Cf. J. N. Behrman, "Notes on the Concept of Capacity and the Measurement of Capital Intensity," Logistics Papers, George Washington University, 1952, pp. 6-16.

mills since the maximum output figure depends upon which of the hundreds of commodities are being produced in the mills. No attempt is made to estimate an average or expected bill of rolled goods or to use an input item as the basis for capacity measurement; hence, no capacity figures are published for this division of the steel industry. Moreover, when the capacity of the steel industry is referred to without qualification, it is the capacity to produce ingots which is implied.[16]

On the questionnaire distributed by the Iron and Steel Institute the instructions for computing the pig iron capacity of blast furnaces and the ingot capacity of Bessemer, open hearth, and electric furnaces are identical. The instructions state that the annual capacity for a firm should be computed by multiplying the number of days in a current year by 24 hours by the number of furnaces, less the number of hours furnaces will be idle because of relining and rebuilding (down time) and holiday shutdowns, times the expected net tons per operating hour.

Lost time does not include time for fettling, repairs to bottoms, banks, and tap holes, hot work repairs, mechanical repairs, and other operating delays. For a shop, the figure computed before allowances for down time is called gross capacity and the figure after down time deductions is called net capacity.[17] From our previous discussion it is seen that gross capacity corresponds to a theoretical capacity, while net capacity is equivalent to a practical capacity.

As clear as these instructions may seem, considerable judgment is required to determine the expected net tons per operating hour for a furnace, whether it be a blast furnace or an ingot furnace. A furnace does not have an unambiguous rated capacity which serves as the basis for the expected net tons per operating hour. For example, the maximum amount of pig iron which can be produced in a blast furnace during some period of time depends upon the quality of the iron ore used and the characteristics of other input materials, especially coke. To account for this variability, the rated capacity of a blast furnace is determined not only by the physical character of the furnace but also by the character of the available materials. Such materials vary greatly in quality. Iron ore in some districts has an iron content of about 60 percent of natural weight, while in other districts the best ore may have only 35 percent metallic content. Hence, the rated capacity of blast furnaces is based upon the operating experiences of each furnace; two people rating the same furnace may easily derive different figures. Moreover, a furnace need not be rated at the same figure over time if operating conditions change, and clearly two furnaces of the same type with the same physical characteristics need not have the same ratings.

Similarly, for a steel-making furnace the rated capacity is based on its performance during a previous full year of continuous operation, and the characteristics of the input again are very important. For an open hearth furnace, for example, about 10 to 12 hours are required to produce each "heat" or output of steel with a mixture of molten pig iron, scrap steel, and limestone. But if some of the scrap is replaced with molten steel from the Bessemer process, as is occasionally done, the "heat" is considerably speeded and the rated capacity would increase.

For new furnaces, both blast and steel-making, the initial capacity rating is based on past experiences of comparable furnaces operating with similar raw materials and under

---

[16]Discussion taken from "Steel Capacity Statistics Based on Performance," Steel Facts, April 1952, pp. 6-7; and Edwin G. Nourse, op. cit., pp. 254-264.

[17]American Iron and Steel Institute, Survey of Steel Capacity, Form 202, January 1, 1954. For a reproduction of the instructions, see Steel Facts, April 1952, p. 6.

similar conditions. These estimates, however, tend to be conservative, and it frequently happens that operating experience reveals that a new furnace has been rated too low.[18]

Estimates tend to be conservative even in the rating of existing equipment, particularly in regard to the effect of improved raw materials and technological advances on potential output rates. Heading the list of things that are increasing ingot capacity is coal washing. In October 1948, U. S. Steel started a huge washer at Gary, West Virginia, that produces 20 to 23 thousand tons of washed coal daily. This coal goes to its plants in the Chicago district, where the improvement is reported to have added at least 500 thousand tons to that area's steelmaking potential, but the full extent of this improvement is not reported in the capacity statistics. Steel officials reportedly feel that such new techniques require study of operating records over a relatively long period before their over-all effect can be calculated. In recent years Republic Steel has been converting to high-top pressure blast furnace blowing. A company official reported that the new technique added 105 thousand tons to blast furnace capacity in 1947 and an additional 75 thousand tons to that figure during 1948, but at the same time he admitted that these calculations are conservative. Alloy and stainless-steel melting time and quality have both been improved by use of oxygen. Reducing time per heat has boosted capacity. For the industry as a whole the gain is substantial, but evidently most shops prefer not to talk about it. Faster charging of open hearths and electric furnaces is another promising field for capacity gains, but apparently little information is available on these improvements.[19]

Some other features of the capacity measures should be noted. The estimates for a given year are computed on the basis of the amount of equipment in the previous year plus all changes in equipment and improvements which will be made before the coming April of the given year. Moreover, all plants which have been idle less than five years are included in the capacity figure for the given year. The officials of the Iron and Steel Institute are aware that this expedient is not entirely satisfactory, since the number of years a plant has been idle does not completely determine its state of repair and potential usefulness. Yet some such arbitrary, but consistent, rule is needed to account for idle equipment. Because estimates tend to be conservative and because investment and improvements for the latter three-quarters of a year are not accounted for in the capacity estimates for a given year, it is easy to see why the operating rate, or utilization ratio, may be over 100 percent. This actually occurred in 1951 when the operating rate for ingot production was 100.9 percent.[20]

Another factor which may permit the operating ratio to go above 100 percent is the allowance for holidays. The present deduction for holidays is four days, even though one or more of these days may actually be worked. In the last few years, for example, when steel has been in short supply, at least three of the holidays were working days. The rigid allowance for holidays thus tends to understate the capacity estimates in some years.

In addition to capacity estimates, operating ratios for the industry are presented in the statistics. The operating ratio is found by dividing yearly production by the annual capacity for the industry. The operating ratio, however, incorporates a lag which was noted previously. In the computation of this ratio for a given year, the production figure is that year's total output, whereas the capacity estimate depends on the previous year's equipment plus all changes and improvements in equipment expected to be completed by April 1st of the given year. In any use of the operating ratio it would probably be useful to correct for this bias.

---

[18] Steel Facts, April 1952, pp. 6-7.
[19] G. F. Sullivan, "Steel Capacity," The Iron Age, January 6, 1949, pp. 198-205.
[20] American Iron and Steel Institute, Annual Statistical Report, 1951, p. 7.

The capacity statistics show that from 1926, when pig iron capacity was first reported, until 1941 the capacity of blast furnaces remained relatively constant, while from 1942 there has been a fairly steady increase in reported capacity. The operating ratio, however, reflects fairly accurately the course of business activity, and appears to dip each time that the level of business falls. The capacity for ingot production, in contrast, increased steadily from 1926 until 1932, remained constant from 1933 to 1937, and finally increased rather steadily from 1938 to 1952, except for minor declines in 1946 and 1947. The growth of ingot capacity has been relatively greater than pig iron capacity, reflecting the increasing importance of scrap steel as part of the charge in Bessemer, open hearth, and electric furnaces. The operating ratio for ingot steel production exhibits substantially the same pattern as does that for pig iron production, except that the former ratio is rather consistently several points higher.[21]

The statistics on capacity for the industry represent a complete coverage of all firms producing pig iron and steel ingots. We have, however, noted some features of the data which indicate that the capacity measures should be treated cautiously and somewhat skeptically. Of prime importance is the fact that rated capacities for furnaces depend heavily upon individual judgments which are not easily subject to verification. We have noted particularly that there is some reluctance to revise capacity figures as innovations occur and operating conditions change. According to one author,[22] capacity estimates generally tend to be conservative and to lag behind improvements in equipment. Other minor deficiencies of the data are that holidays may be worked, even though deductions are allowed for them, and that in the operating ratio formula there is a lag between the capacity data and the production figures. Like much statistical information, it is not an easy matter to determine what degree of specious or irresponsible reporting occurs in the steel industry, and I am at a loss to make an estimate of this factor.

2. Petroleum Refining

For the petroleum industry, statistics on operations are largely compiled by the Bureau of Mines and the American Petroleum Institute. The Bureau of Mines publishes its findings on capacity estimates yearly in an Information Circular, generally appearing in the latter half of the given year, which purports to show the capacity as of January 1 of that year.[23] Reports have appeared annually from 1918, except for the war years of 1942 through 1945. The American Petroleum Institute collects some data originally from the refineries, but its estimates of refinery capacity are obtained from the Bureau of Mines. The Institute publishes its findings in the annual Petroleum Facts and Figures and the weekly Statistical Bulletin. The Oil and Gas Journal also presents an annual survey for petroleum refining, and for the years 1942 through 1945 it is the only original source for capacity data.[24]

In the Bureau of Mines annual survey, capacity figures are estimated independently for the two major divisions of the refinery: the distillation process and the cracking process. The former is reported as refinery capacity and the latter as cracking plant capacity. For each category an estimate is made for the plants actually operated during the year, for shutdown

---

[21] American Iron and Steel Institute, Annual Statistical Report, 1952, pp. 27 and 37.

[22] G. F. Sullivan, op. cit., pp. 202-203.

[23] For example, cf. F. S. Lott and J. G. Kirby, "Petroleum Refineries, Including Cracking Plants, in the United States, January 1, 1950," Information Circular 7578, Bureau of Mines, August 1950, 19 pp.

[24] For example, cf. R. B. Tuttle, "U. S. Refineries Have Crude-Oil Input Capacity of 6.75 Million Barrels," Oil and Gas Journal, March 23, 1950, pp. 302-304.

plants, and for plants in the process of being built during the year. Shutdown capacity includes capacity of inoperative portions of operating plants, as well as capacity of plants completely shut down. Capacity of facilities under construction includes the capacity of construction to augment existing plants, as well as new plants. Before 1938, however, shutdown capacity includes only capacity of plants completely shut down, and capacity of facilities under construction includes only that of new plants in process of erection. In each circular a time series is presented for total refinery plant capacity and cracking plant capacity, according to the threefold classification of operating, shutdown, and building. For the given year summaries are presented for refineries and cracking plants by districts and by states, and a summary is presented for cracking plants by types, i.e., reformed, thermal, and catalytic. A detailed report is also offered for each company in the United States where these are classified according to the type of basic atmospheric distillation equipment employed.

The distillation capacity of a refinery is represented in terms of the input, crude oil. According to this concept, the capacity of a plant is the maximum amount of crude oil which can be processed in an average day when the equipment is in complete operation. An average daily figure is reported, since the capacity figure purportedly shows the maximum daily crude runs to stills that can be maintained for an extended period. By this is meant that the daily figure allows for necessary shutdown time for routine maintenance, repairs, and breakdowns, but assumes no interruptions from strikes, a continuous supply of crude, no material change in operating methods, and a demand sufficient to take the entire production. The reporting procedure evidently is not as standardized as for the steel industry, since the amount of crude oil intake to be deducted for necessary interruptions to continuous refining depends entirely upon the individuals estimating plant capacity. Apart from allowances for interruptions, the rating of a plant also depends heavily upon individual judgment since the estimate depends upon the quality of the crude oil distilled, which may vary considerably, and upon the previous operating experience for the plant, which is dependent not only on the size and the quality of the distillation equipment but also on the efficiency of the working force and the auxiliary equipment, previous price policies, etc.

The cracking plant capacity is computed in a similar manner and is subject to similar difficulties. But whereas distillation capacity is derived in terms of an input, cracking plant capacity is computed as the amount of finished product in barrels, primarily gasoline, which can be produced in an average day by using charging stocks of the type in current use.[25]

For the distillation process, capacity is not measured in terms of output, primarily because of a product mix problem. The difficulty arises because of the large number of products which are feasibly producible from crude oil by the combination of distillation and chemical treatment; and since many products have a seasonal demand, the product mix for a refinery varies over the year, even if it specializes in certain commodities. As we noted previously, however, the input measure appears to be a reasonably good compromise for these difficulties.

The capacity statistics for the petroleum industry indicate that total distillation capacity has increased uninterruptedly, except for minor setbacks in 1933 and 1943, from 1918, the date of the first annual Bureau of Mines survey, until the present. From 1921, the year operating capacity was first reported, it has grown each year, except for minor declines for five

[25] The information on capacity, except where otherwise noted, is taken from the Bureau of Mines Information Circulars, various issues, and Petroleum Facts and Figures, various issues.

421661 O - 57 - 2

years; but, as is to be expected, the rate of growth is less steady than for total installed distillation capacity. Until 1943, the capacity of shutdown plants averaged roughly 10 percent of total capacity, but with fluctuations between 4 and 14 percent. Since 1943, the capacity of shutdown plants has decreased fairly steadily and in 1953 was only 2 percent of total capacity.

The capacity of cracking plants was first reported in 1925 and until 1938 was computed in terms of the charging stock processed by the plants. As nearly as can be determined from comparing the two bases of measurement, total cracking plant capacity has grown steadily without interruption since 1925, and since 1938 appears to have approximately tripled. During this period the relative amount of shutdown capacity has fluctuated from 4 percent to around 22 percent of total installed capacity. For every year since 1939, however, the capacity of shutdown plants has been less than 10 percent of total installed capacity, decreasing to 1.2 percent in 1953.

No time series on operating ratios are presented, but the over-all operating ratio for distillation plants for the given year is indicated in each Bureau of Mines annual survey. For years in which business activity is at a high level the operating ratio between production and total installed capacity appears to be around 90 percent.[26]

In the petroleum industry a concept of capacity and interest in measurements of capacity do not seem to play as central a role as in the steel industry. In the trade journals, e.g., Petroleum Processing, Petroleum Refining, and the Oil and Gas Journal, one rarely finds any discussion of the measurement of capacity. Generally, any reference made to capacity is to report, uncritically, the findings of the Bureau of Mines. In its annual survey the Bureau of Mines reports briefly and generally the capacity concept which is the basis of measurement for the industry, but it does not discuss either the reporting procedure or the accuracy of the capacity estimates. In an interview with officials of a petroleum refining firm, however, the author was warned that the capacity measures reported in the various publications must be used with caution. It was stated that one cannot place much credence in the accuracy of the measures of capacity since the estimates tend to be biased fairly consistently on the optimistic side. For many companies it would take major investments for restoring and improving the plant before the reported capacity figure could be approached, and apparently some of the smaller refineries outrightly and without justification tend to overstate the capacities of their plants.

3. The Electric Utility Industry

For the electric utility industry, statistics on the production of electric power and on the capacity of generating plants in the United States are reported systematically by the Federal Power Commission and the Edison Electric Institute.[27] The data presented by the Federal Power Commission is based on required monthly reports to the commission; and the coverage, which is substantially 100 percent, comprises around 3700 generating plants operated by some 1400 electric utilities. Included are plants of both the privately owned electric utilities and the publicly owned group. Data on electric energy generated by industrial firms and agricultural organizations for their own use are not included in comprehensive form. The capacity of small standby plants, operating infrequently and having little or no production, is included in the

---

[26] J. G. Kirby, "Petroleum Refineries, Including Cracking Plants, in the United States, January 1, 1953," Information Circular 7667, Bureau of Mines, September 1953.

[27] Federal Power Commission, Electric Power Statistics, and Edison Electric Institute, Statistical Bulletin.

summary data, though regular monthly reports are not requested from such plants.[28] The Edison Electric Institute data are derived partly from their own sources and partly from the official Federal Power Commission information.

For the industry various measures of capacity are computed for different purposes. The basic data for each measure, however, is the rated capacity, or name-plate rating, of generating equipment.[29] In brief, the name-plate rating is the manufacturer's estimate of the number of kilowatts which can be produced in one hour by the generator, but caution is required in accepting the rating as an accurate estimate of the productive ability of the equipment.

The name-plate rating is the basis of guarantee from the manufacturer to the purchaser covering the efficiency of a machine under certain specified conditions. In practice, because of differences in operating conditions or of modifications since installation, the actual productive ability of generating units may be quite different from the name-plate rating. In their own reports most companies carry a "capability rating" for equipment, instead of the name-plate rating, since it more nearly sets forth kilowatt power under actual operating conditions.

Neither the name-plate nor the capability rating, however, gives the full measure of power of generating facilities. Most generating equipment driven by steam turbines or by steam engines is capable of considerable overload, often for quite an extended period. With generators driven by water wheels or water turbines, stream flow or water flow govern the possible output, but even under adverse conditions such generators are capable of carrying a substantial overload.[30]

With these limitations in mind we may consider the different capacity measures. The estimate given as the industry capacity is simply the sum of the name-plate ratings of the equipment in service at the end of the year, as reported to the Federal Power Commission. To compute annual utilization ratios (commonly called plant factors), an annual capacity is obtained for the system by multiplying the name-plate ratings of the equipment in service at the end of the preceding year plus one half of the net increase during the year in question by 8760, the number of hours in a year. This measure thus corresponds to a theoretical capacity where potential output is considered as 100-percent use of the installed equipment operating continuously throughout the year.

In comparing peak loads (the highest hourly rate of generation of electric power) with capacity, two additional capacity measures are used: dependable capacity and net assured capacity. Dependable capacity is defined as the name-plate ratings of all generating plants that can be relied upon for active or standby service, plus the capability of hydro plants under the most adverse water conditions and plus net firm purchases—all at the time of the annual system peak. Net assured capacity is dependable capacity less reported required equipment reserves.[31] These two concepts, especially the latter, correspond roughly to practical capacities. In net assured capacity no explicit allowances are made for maintenance and repair time, but the deduction for equipment reserves apparently accounts for this category. An electric generator reportedly requires very little repair, but since electric utilities are required to meet any demands for power made on the system by users, it is necessary to keep sufficient

---

[28]Federal Power Commission, "Production of Electric Energy and Capacity of Generating Plants 1954," Electric Power Statistics, Binder Section 1, 1955, p. 1.

[29]Capital equipment in the industry is divided into three major classes: generation, transmission, and distribution. On the average, for the individual firm there appears to be about equal investment in generation and distribution, while investment in transmission accounts for somewhat more than half of either of the former categories.

[30]Edison Electric Institute, Statistical Bulletin, 1952, pp. 4-5.

[31]Federal Power Commission, op. cit., Report No. 55-66, p. 16.

reserve equipment to forestall any power shortage caused by breakdowns of machinery. According to the data, reported required reserves presently average around 8 percent of installed name-plate capacity.[32]

No information is currently given concerning the policies used to determine required reserves, but Nourse maintained that the reserve cannot safely be less than the capacity of the largest generating unit or the largest tie-in in the system. He also reported that more conservative managers maintain a reserve equal to the largest two units, visualizing the chance that some day at the hour of the peak load, when the largest unit is down for repairs, the second largest also may suddenly fail.[33] The stream flow limitation on potential output is becoming less and less important in the United States, primarily because hydro plants are continually becoming percentagewise less important in the industry but also because it is becoming common to store water power.

Once the annual capacity is given, the plant factor is computed by dividing annual output in kilowatts by the capacity. The average annual plant factor for all plants is presently around 60 percent, though there exists considerable variation in the system. The main reasons for this relatively low ratio are easily seen to be the theoretical capacity measure used as the divisor and the seasonal demand for electric energy. The capacity concept does not allow for the need to withdraw generating equipment from service for periodic overhauling or repair, for equipment held in reserve, and for the variation in stream flow for hydro plants. As far as the seasonal demand is concerned, the system load varies hourly, as well as daily and monthly, but enough equipment must be ready for use to meet peak demands.

It is interesting historically that from 1920 through 1938 the annual plant factor for the system averaged around 36 percent, except for the depression years of 1930-1935. From 1938, however, there has been an upward secular trend—the ratio has increased steadily from 35.5 percent in 1938 to 58.2 in 1953.[34] According to the Edison Electric Institute, this improvement is the result of several things. The two most important are: (1) the greater reliability of generating and transmission equipment which, with integration and interconnection between electric systems, has reduced the spare equipment required for emergency or reserve; and (2) the constant expansion of uses for electrical devices has prolonged and made more uniform the demand for electricity throughout the day. Data on the plant factor extend back to the Census of 1902 when the ratio was 25.7 percent and are available at five-year intervals until, beginning with 1926, we have the annual record of the Edison Electric Institute.

The increase in the plant factor since 1902 would appear still larger were it possible to allow for changes in the practice of rating electrical equipment. In the early days of the industry, the typical manufacturer, held to a guarantee of performance and with little experience to guide him, seems to have rated his generators more conservatively than today. Users finding their machines capable of operating safely at loads greater than name-plate ratings operated them at loads above that contemplated by the designers. If plant factors for the early years were calculated on the basis of safe capacities, the factors would be somewhat lower than reported.[35]

To measure the variations in demand, the electrical industry has developed the concept of the annual load factor. This ratio should not be confused with the plant factor, though it is

---

[32] Ibid., p. 6.
[33] Edwin G. Nourse, op. cit., pp. 328-331.
[34] Federal Power Commission, op. cit., p. xiv.
[35] Edwin G. Nourse, op. cit., pp. 319-321, and pp. 478-480.

computed in a similar manner. For a plant the load factor is the total annual kilowatts produced, divided by the peak load on the plant (the highest hourly rate of electrical energy generated by the plant for the year), times 8760, the number of hours in a year. In other words, it is the average hourly production divided by the peak load. For the system the peak load is computed as the sum of the peak loads for all the plants, and it is a noncoincident peak since the individual peaks may have occurred at different hours, on different days, and in different weeks. A true peak, determined simultaneously for the whole system for the same hour, would be substantially less than the noncoincident peak.[36] The similarity of the noncoincident load factor to the plant factor is obvious, the difference being that in the former the divisor is the peak load times 8760 while in the latter it is the annual capacity.

As to the recent accuracy of the data, some evasive reporting may occur or some omissions may be present, but the reporting procedures and the characteristics of the equipment appear to insure that the margin of error is very small, at least as far as the reporting of name-plate capacity is concerned. Each piece of generating equipment has a manufacturer's name-plate rating in terms of kilowatts, and each plant is required to report these ratings monthly to the Federal Power Commission. As noted previously, however, even though we may feel confident that the data reported are accurate, caution is required in accepting the name-plate capacity as an accurate rating for generating equipment.

THE USES OF INDUSTRY MEASURES OF CAPACITY

The uses of industry measures of capacity, of course, are highly conditioned by the accuracy of the observations. We have already argued that on this matter the user of capacity data must be extremely cautious, particularly since, in many cases, it is very difficult to obtain information on the accuracy of the estimates. We need not repeat our previous warnings in this regard; instead, we shall confine the discussion to uses of the measures apart from the limitations imposed by inaccurate observations.

Capacity measures of this type are, mainly, a function of the existing level of capital equipment in physical units and the production relation between services from the capital equipment and the commodities produced by the industry. Whether or not there is an active interest in an industry in actual capacity measures, the information required for the derivation of capacity estimates is immediately important for the planning of production and investment by industry entrepreneurs. Other variables which may enter in the determination of capacity, e.g., down time and other repair time, also are of importance in this respect. A firm may not explicitly combine such data to derive capacity estimates, but implicitly, at least, capacity figures of some types are important for the rational functioning of the organization. Consequently, whether for understanding the formulation of plans or for making predictions, capacity measures for industries, where they are obtainable, potentially, at least, embody relevant information.

Generally speaking, it is clear that industry measures of capacity contain useful information. The problem now is what specific uses do they have for economic analyses. Conceivably, the estimates may be useful as measures of the ability of industry to produce output during the present period of time, as measures of the growth pattern of industries, or as measures of capital equipment which may be incorporated into investment models.

---

[36] Edison Electric Institute, Statistical Bulletin, 1952, p. 37.

In the first instance, as indicators of the ability to produce for an industry, the measures have obvious limitations. The usual industry capacity estimate purports to show the productive ability of existing physical equipment, but this ability, of course, is dependent upon the full set of conditions occasioning the use of the equipment, and the measure holds only for the special conditions which are the basis of measurement. Unfortunately, the majority of reported industry capacities are computed on the assumptions that the costs of using the equipment are not relevant, that there is a free supply of complementary inputs, and that there is sufficient demand to purchase any quantity of the commodities produced. These usual assumptions do not realistically portray existing conditions in an industry and are generally not the type of conditions which would interest the investigator of the industry's ability to produce. For these reasons industry measures of capacity, by themselves, are not very useful for this purpose.[37]

As indicators of growth, industry measures of capacity are on a sounder basis if they are estimated consistently from year to year or if appropriate corrections can be made for any inconsistencies which do appear in the data. They are frequently used for this purpose in trade journals and occasionally in studies on the growth pattern of industries.[38] In this use they may have certain advantages over other measures of growth, such as the value of sales or output, balance sheet totals, or the value of durable production equipment, in that they allow both for changes in the amount of durable production equipment in the industry and for technological changes, whether these relate to improvements in existing equipment or to the introduction of more efficient machinery. Moreover, they may have additional advantages over value measures of growth in that they enable the user to show the growth pattern of the industry without being troubled by valuation problems. Compared with output time series, their advantage is that they are less apt to be influenced by cyclical fluctuations in business activity and consequently they more readily indicate trend factors which tend to be obscured in the output series. On this latter point, Hickman showed, by plotting capacity series, that changes in capacity over time seem to be largely independent of cycles in production. The general impression seems to be one of secular growth or decline, and this is confirmed when the data are studied systematically.[39] An obvious reason would appear to be that the firm does not vary the quantity of equipment it holds with every short-run fluctuation in demand; this would be too costly. Instead, it will tend to vary the intensity with which the machinery is used for short periods. Another factor would appear to be that technological advances occur even in depressed years and tend to raise the capacity rating for existing equipment as well as for new equipment. The major drawbacks of the capacity measures, in this use, are that the estimates do not indicate the age distribution of machinery and may include an unspecified amount of output which is attributable to obsolete or long-idle machinery. While capacity measures tend to be upgraded for technological advances, it is difficult to account for loss in efficiency of aged or obsolete machinery in noncost measures. These limitations may partially account for the cyclical insensitivity of capacity estimates. During depressed years firms may continue to carry obsolete equipment and may not replace operating equipment as frequently as in prosperous years but, nevertheless, may not downgrade the capacity estimates correspondingly.

---

[37] For a discussion of more appropriate capacity measures for this use, cf. Edward Zabel, loc. cit., pp. 237-243.

[38] One example is the study by Bert G. Hickman, "Cyclical Variations in Capacity and Capacity Utilization in Selected Industries," Conference on Research in Income and Wealth, National Bureau of Economic Research, Inc., October 9-11, 1953.

[39] Bert G. Hickman, op. cit., pp. 12-13.

Another drawback in using capacity series as indicators of growth is that it is dangerous to make direct comparisons among different industries as to absolute changes in growth on the basis of capacity estimates. The obvious reason is that different industries use different concepts of capacity. And this difference in concepts leads to estimates of potential output which are not directly comparable.

As variables in investment models, both industry capacity estimates and utilization ratios may be useful. A capacity estimate may be used as a measure of the existing level of capital equipment in an industry, which can be easily related to an output or demand variable and which, again, avoids valuation problems. In conjunction with utilization ratios, it is possible to set up an investment model which attempts to predict changes in capacity rather than directly predict changes in capital equipment. The former type of model would have the advantages that it avoids some valuation problems relating to capital equipment, the measure of capital equipment is more readily related to production, and predicting changes in capacity may be simpler than predicting changes in physical equipment in the sense that variations in capacity include both technological advances and changes in the amount of equipment. On the latter point, for example, firms may meet increases in demand by improving existing equipment rather than by adding new equipment. Models which predict variations in capacity would account for this, whereas the usual model relating directly to capital equipment generally does not allow for this type of adjustment. If the ultimate problem, however, is to predict changes in capital equipment, the advantage of the capacity model may be nullified in the sense that, in relating changes in capacity to equipment, the problems which have been overcome in the model arise again in the final transformation.

Utilization ratios may be helpful in predicting the amount by which capacity increases in response to an increase in demand but, as with the capacity estimates, these ratios should be used with care. The user must be cautious about making inferences from these data, particularly in regard to the absolute levels of the ratios. For example, if industry A on the average has a utilization ratio of 60 percent and industry B 95 percent, one cannot infer that equipment is being used less efficiently or that more idle or obsolete equipment exists in industry B. The difference may be entirely due to different capacity concepts, the accuracy of the data, and different operating characteristics for the two industries. Industry A may use a "theoretical" capacity concept as the basis of measurement and may include all idle equipment, whereas industry B may use a "practical" capacity concept and include only equipment which is being currently used or which has been idle for less than, say, three years. In industry A the capacity rating of equipment may depend on judgments which would vary among individual raters, but in industry B the ratings may rely on verifiable engineering data. Also, industry A may be subject to seasonal demand, while in industry B demand may be fairly constant over the year.

Even for the same industry the user must be cautioned when making inferences from year-by-year comparisons of utilization ratios. In discussing capacity estimates, we have already noted that over the years the reporting procedures may change in an industry or the extent of bias in the estimates may change (e.g., for the electric power industry); these anomalies, of course, will affect the utilization ratios and should be guarded against. We have also noted that in the case of electric power there has been an upward trend in the utilization ratio because of improvements in equipment, the practice of interconnection of power systems, and some leveling out of seasonal demand. When comparing utilization ratios for the same industry, particularly over long periods, the user must be alert for such trends. In comparison with

previous years, a relatively low utilization ratio may also reflect only the fact that the industry is increasing capacity in anticipation of future demand, a healthy sign, rather than validating an inference that equipment is being less efficiently employed than in previous years or that the industry is relatively depressed.

In summary, the discussion in this study has revealed the range of concepts which are the bases for estimates of industry capacity and has indicated some possible uses for the data. To date, however, industry measures of capacity have found relatively little employment in economic analyses.

* * *

A BOUND ON THE USE OF INEFFICIENT INDIVISIBLE UNITS

Martin J. Beckmann
Center for Advanced Study in the Behavioral Sciences
and
Jack Laderman
Office of Naval Research

The authors discuss a problem of resource allocation in which the usual economic assumption of perfect divisibility of factors does not apply.

The existing body of economic theory on resource allocation rests rather critically on the assumption of perfect divisibility of factors. Whenever the factors involved are indivisible and are to be combined in small numbers, the principles of marginal analysis offer no guidance, and entirely new considerations seem to be called for. To obtain an idea of the peculiar difficulties that may then be faced, we propose to study the following simple but perhaps not unrepresentative problem.

Suppose that a certain number of passengers is to be transported on a given route for which planes of two types are available. Type 1 planes have larger capacity, s_1, but smaller operating cost, c_1, per seat. Provided the total operating cost of the type 2 planes is smaller,

$$s_2 c_2 < s_1 c_1 , \tag{1}$$

the cheapest transportation for a certain number of passengers may call for the use of one or more of the less-efficient type 2 planes. In this paper we obtain a bound on the number of type 2 planes that should ever be used. This bound is of interest in itself for the question of the maximal feasible investment in small planes. But its main usefulness is in reducing the number of trial and error solutions that need be examined in determining the optimal combination of planes for transporting a given number of passengers. Computational problems proper, however, will not be considered here.

Suppose the optimal solution calls for the use of n type 2 planes, with possibly some number of type 1 planes. It then follows that ns_2 is not divisible by s_1. This is so because if $ns_2 = ks_1$ (k an integer), then the passengers in the n type 2 planes could be transported more economically in k type 1 planes.

*Manuscript received April 27, 1956.

421661 O - 57 - 3

Let

$$\frac{ns_2}{s_1} = m + \frac{r}{s_1}, \tag{2}$$

where m is an integer and $1 \le r \le s_1 - 1$.

We know that the number of passengers in the n type 2 planes can not be transported more economically with the use of one or more type 1 planes. In particular, there would be no saving in using $(m + 1)$ type 1 planes. Hence

$$n s_2 c_2 \le (m + 1)\, s_1 c_1 .$$

Then by use of (2) we have

$$ns_2 c_2 \le \left(\frac{ns_2}{s_1} - \frac{r}{s_1} + 1\right) s_1 c_1$$

$$\le \left(\frac{ns_2}{s_1} - \frac{1}{s_1} + 1\right) s_1 c_1$$

$$\le ns_2 c_1 - c_1 + s_1 c_1 ;$$

and finally,

$$n \le \frac{s_1 - 1}{s_2} \cdot \frac{c_1}{c_2 - c_1} . \tag{3}$$

Similarly, there would be no saving in using m type 1 planes to transport ms_1 passengers and using type 2 planes to transport the remaining passengers. Since the number of remaining passengers is at most $(ns_2 - ms_1)$, the number of type 2 planes needed is not greater than

$$\frac{ns_2 - ms_1}{s_2} + \frac{s_2 - 1}{s_2} .$$

It follows that

$$ns_2 c_2 \le ms_1 c_1 + \left(\frac{ns_2 - ms_1}{s_2} + \frac{s_2 - 1}{s_2}\right) s_2 c_2 ,$$

$$\le ms_1 c_1 + ns_2 c_2 - ms_1 c_2 + (s_2 - 1)\, c_2 .$$

We then get

$$ms_1 \leq \frac{(s_2 - 1)\, c_2}{c_2 - c_1}.$$

By use of (2) we have

$$ns_2 - r \leq \frac{(s_2 - 1)\, c_2}{c_2 - c_1},$$

or

$$n \leq \frac{s_2 - 1}{s_2} \cdot \frac{c_2}{c_2 - c_1} + \frac{r}{s_2},$$

$$\leq \frac{s_2 - 1}{s_2} \cdot \frac{c_2}{c_2 - c_1} + \frac{s_1 - 1}{s_2}.$$

Then

$$n < \frac{c_2}{c_2 - c_1} + \frac{s_1 - 1}{s_2}. \tag{4}$$

Combining (3) and (4), we have

$$n \leq \text{Min}\left(\frac{s_1 - 1}{s_2} \cdot \frac{c_1}{c_2 - c_1}, \frac{s_1 - 1}{s_2} + \frac{c_2}{c_2 - c_1}\right). \tag{5}$$

For brevity, write

$$\lambda = \frac{s_1 - 1}{s_2}, \quad \rho = \frac{c_1}{c_2}.$$

Then (5) becomes

$$n \leq \text{Min}\left(\frac{\rho}{1 - \rho}\lambda, \frac{1}{1 - \rho} + \lambda\right). \tag{6}$$

From (1) we have a lower bound on ρ, namely

$$\rho > \frac{s_2}{s_1}. \tag{7}$$

The first term of the right hand side of (6) is the smaller one when $\rho < \frac{1}{2} + \frac{1}{2\lambda}$, and the second term is the smaller one for larger values of ρ. The second term increases with ρ and becomes $(s_1 - 1)$ when $\rho = 1 - \frac{1}{\lambda(s_2 - 1)}$. Since $(s_1 - 1)$ is an obvious bound for n, (6) should not be used for larger values of ρ. Also, since n is an integer, we can omit any fractional part of the number obtained for its bound. Letting [x] stand for the largest integer not exceeding x, we can summarize the above by

$$
\begin{aligned}
n &\le \left[\frac{\rho}{1-\rho}\lambda\right], && \frac{s_2}{s_1} < \rho \le \frac{1}{2} + \frac{1}{2\lambda}, \\
&\le \left[\frac{1}{1-\rho} + \lambda\right], && \frac{1}{2} + \frac{1}{2\lambda} \le \rho \le 1 - \frac{1}{\lambda(s_2 - 1)}, \\
&\le s_1 - 1, && \rho > 1 - \frac{1}{\lambda(s_2 - 1)}.
\end{aligned} \tag{8}
$$

The bounds (8) increase with both ρ (the ratio of costs) and λ (the approximate ratio of sizes), as must be expected for the true maximum of n.

As an example, suppose the capacities of DC-6's and DC-3's on first class flights are 58 and 36 passengers, respectively. Then

$$
\lambda = \frac{19}{12}, \quad \frac{s_2}{s_1} = \frac{18}{29}, \quad \frac{1}{2} + \frac{1}{2\lambda} = \frac{31}{38}, \quad 1 - \frac{1}{\lambda(s_2 - 1)} = \frac{653}{665}.
$$

The bounds on the number of DC-3's which need be considered for possible use, for various cost ratios, are given below.

ρ	$\frac{12}{19}$	$\frac{13}{19}$	$\frac{14}{19}$	$\frac{15}{19}$	$\frac{31}{38}$	$\frac{16}{19}$	$\frac{17}{19}$	$\frac{18}{19}$	$\frac{653}{665}$
Bound for n	2	3	4	5	7	7	11	20	57

The problem of finding the optimal number of type 1 and type 2 planes to transport a given number of passengers can be formulated as a discrete linear programming problem.

Let

x_1 = number of type 1 planes,

x_2 = number of type 2 planes,

p = number of passengers to be transported.

Then, we seek integral values of x_1 and x_2 which minimize

$$x_1 s_1 c_1 + x_2 s_2 c_2 , \tag{9}$$

subject to the constraints,

$$x_1 s_1 + x_2 s_2 \geq p \tag{10}$$

$$x_i \geq 0 . \tag{11}$$

Denote the optimal solution by x_1^0, x_2^0. Comparing its cost with that of using

$$x_1 = x_1^0 + \left[\frac{x_2^0 s_2 + s_1 - 1}{s_1}\right]$$

$$x_2 = 0 , \tag{12}$$

one can obtain inequality (3) with x_2^0 for n.

Also, letting

$$x_1 = x_1^0 + \left[\frac{x_2^0 s_2}{s_1}\right]$$

$$x_2 = \left[\frac{p - x_1 s_1 + s_2 - 1}{s_2}\right], \tag{13}$$

one can get inequality (4).

The linear programming formulation permits the following geometric interpretation of the allocation problem. In Figure 1, combinations of numbers of the two types of planes are represented by points of a quadratic lattice having integral coordinates. All combinations having a capacity for at least p passengers are located on and above a line, to be called an isocapacity, whose slope is $- s_1/s_2$ and whose intercept on the x_i axis is p/s_i. The family of isocosts (lines of constant cost) consists of lines of greater slope, $- c_1 s_1/c_2 s_2$, than the isocapacities. The optimum combination lies on that isocost nearest to the origin which passes through a lattice point to the right of or on the isocapacity line for

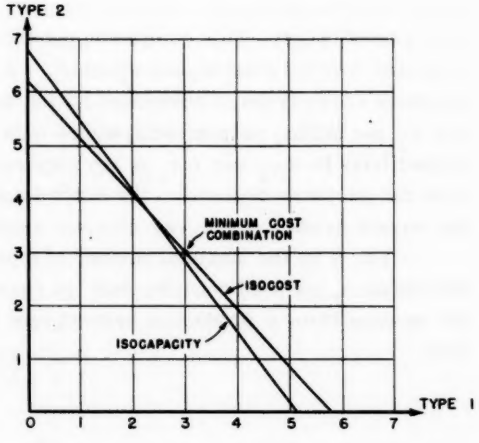

Figure 1

p passengers. This means that the interior of the triangle, formed by the isocost on the right, the isocapacity on the left, and the x_1 axis at the bottom, should not contain any lattice points.

If there are more than two unit sizes of the indivisible factor, the inequalities (8) may still be applied to yield a bound on the input of any unit in terms of the sizes and costs of larger units. Let the units be labelled in order of decreasing size and (hence) of increasing average costs,

$$s_1 > s_2 > \cdots\cdots > s_n ,$$

$$c_1 < c_2 < \cdots\cdots < c_n ,$$

with

$$s_1c_1 > s_2c_2 > \cdots\cdots > s_nc_n .$$

Define

$$\lambda_{ij} = \frac{s_i - 1}{s_j} , \qquad \rho_{ij} = \frac{c_i}{c_j} ,$$

then (8) becomes

$$n_j \leq \underset{i<j}{\text{Min}}\ \text{Min}\left(\left[\frac{\rho_{ij}}{1 - \rho_{ij}} \lambda_{ij} \right] , \left[\frac{1}{1 - \rho_{ij}} + \lambda_{ij} \right] , s_i - 1 \right) . \tag{14}$$

The constraint $i < j$ is essential, for we must have $\rho_{ij} < 1$.

Every cost minimization problem suggests an associated maximization problem of similar structure. (This should not be confused with the duality principle of linear programming). For example, in the airplane problem one can ask for a bound on the number of type 2 planes that should ever be used to transport the maximum number of passengers having available a fixed budget. This companion piece will be treated by specific reference to a more practical revenue maximization problem. A firm with a limited capital budget considers the purchase of two types of equipment having unit costs of q_1 and q_2 and yielding returns of r_1 and r_2 per dollar, respectively, where $q_1 > q_2$ and $r_1 > r_2$. (Here q_i and r_i correspond respectively to s_ic_i and $1/c_i$ in the airplane problem). Then the best combination that the firm can purchase depends on the exact amount of capital available. But we may ask again for the largest number of the less-efficient production units that should ever be bought.

Let n be the maximal number of type 2 production units that should ever be bought. To determine n, we need consider only the case that the budget limit q equals nq_2. Comparing the revenue from n production units of type 2 with that from the use of only type 1 units, we have

$$nq_2r_2 \geq \left[\frac{nq_2}{q_1}\right] q_1r_1 ,$$

$$nq_2r_2 \geq \left(\frac{nq_2}{q_1} - \frac{q_1 - 1}{q_1}\right) q_1r_1 ,$$

$$nq_2r_2 \geq nq_2r_1 - (q_1 - 1)\, r_1 ,$$

$$n \leq \frac{q_1 - 1}{q_2} \frac{r_1}{r_1 - r_2}, \tag{15}$$

which is analogous to (3).

Next compare the revenue from n production units of type 2 with that from a combination of m units of type 1 and as many units of type 2 as the remaining budget permits. We have

$$mq_1r_1 + \left[\frac{nq_2 - mq_1}{q_2}\right] q_2r_2 \leq nq_2r_2 .$$

Since

$$\left[\frac{nq_2 - mq_1}{q_2}\right] \geq \frac{nq_2 - mq_1}{q_2} - \frac{q_2 - 1}{q_2} ,$$

it follows that

$$mq_1r_1 + nq_2r_2 - mq_1r_2 - (q_2 - 1)\, r_2 \leq nq_2r_2 ,$$

and then

$$m \leq \frac{q_2 - 1}{q_1} \frac{r_2}{r_1 - r_2}. \tag{16}$$

Now let m take on its largest value compatible with a given $q = nq_2$. Then

$$nq_2 \leq (m + 1)\, q_1 - 1 ,$$

and from (16) we get

$$nq_2 \leq \left(\frac{q_2 - 1}{q_1} \frac{r_2}{r_1 - r_2} + 1\right) q_1 - 1 ,$$

$$n \leq \frac{q_2 - 1}{q_2} \frac{r_2}{r_1 - r_2} + \frac{q_1 - 1}{q_2} .$$

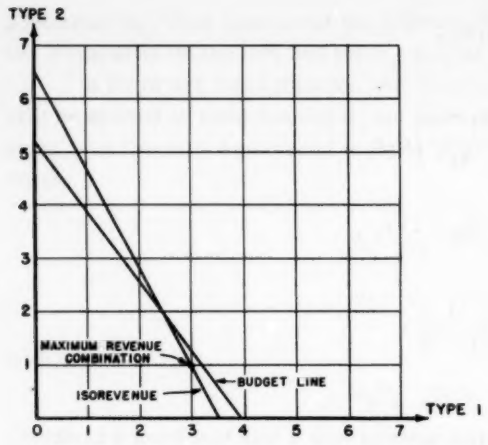

Figure 2

Finally,

$$(17) \qquad n < \frac{r_2}{r_1 - r_2} + \frac{q_1 - 1}{q_2},$$

which corresponds to (4). As in the airplane problem, here also there is an obvious upper bound of $(q_1 - 1)$ for n.

The geometric interpretation of the solution is similar to the one in the airplane problem. In Figure 2 all combinations which cost less than the fixed budget are to the left of the budget line. We let isorevenue lines denote combinations of constant revenue. Then the optimal combination lies on that isorevenue line which passes through a lattice point to the left of or on the budget line and leaves no lattice point in the interior of the triangle bordered by the isorevenue line on the left, the budget line on the right, and the x_1 axis at the bottom.

If there are more than two types of units, we let

$$\phi_{ij} = \frac{q_i - 1}{q_j}, \quad \psi_{ij} = \frac{r_i}{r_j},$$

and a bound on the number of units of the j-th type that should ever be purchased is given by

$$(18) \qquad n_j \leq \operatorname*{Min}_{i<j} \operatorname{Min}\left(\left[\frac{\psi_{ij}}{\psi_{ij} - 1}\,\phi_{ij}\right], \left[\frac{1}{\psi_{ij} - 1} + \phi_{ij}\right], q_i - 1\right),$$

which corresponds to (14).

One important difference between the results in these two problems is that in the maximization problem the bound given by (15) should be ignored because it is never less than the one in (17). Hence (18) can be replaced by

$$(19) \qquad n_j \leq \operatorname*{Min}_{i<j} \operatorname{Min}\left(\left[\frac{1}{\psi_{ij} - 1} + \phi_{ij}\right], q_i - 1\right).$$

* * *

VARIANTS OF THE HUNGARIAN METHOD FOR ASSIGNMENT PROBLEMS[1]

H. W. Kuhn
Bryn Mawr College

The author presents a geometrical model which illuminates variants of the Hungarian method for the solution of the assignment problem.

1. INTRODUCTION

The Hungarian Method [1] is an algorithm for solving assignment problems that is based on the work of D. König and J. Egerváry. In one possible interpretation, an assignment problem asks for the best assignment of a set of persons to a set of jobs, where the feasible assignments are ranked by the total scores or ratings of the workers in the jobs to which they are assigned. It is a special case of the transportation problem, which asks for the allocation of a homogeneous commodity from given supplies at a set of sources to satisfy prescribed demands at a set of destinations, while minimizing the total source-destination transportation costs. The extension of the Hungarian Method to transportation problems can be found in the work of L. R. Ford, Jr., and D. R. Fulkerson [2] and of J. Munkres [6]. Comprehensive reviews of these problems may be found in the papers of M. M. Flood [3] and T. S. Motzkin [4].

The original purpose of this note was to present a modification of the Hungarian Method, inspired by an algorithm of Marshall Hall, Jr. [5], for the choice of a system of distinct representatives for a family of sets. But is now seems that a synthesis of all of the available variants would be more valuable. Accordingly, the paper now presents a geometrical model that is adequate to define and compare the algorithms contained in [1], [2], [5], and [6]. In Section 2, the Hungarian Method is reviewed in outline so as to isolate the procedure that is altered in these variants. The modification suggested by Hall's work is presented in Section 3. A translation into graph-theoretical terms of the problem solved by all four versions is the subject of Section 4; this section is self-contained and can be read by itself. A final section compares these alternative solutions by means of this geometric model.

2. RESTATEMENT OF THE HUNGARIAN METHOD

As considered in this paper, the assignment problem asks: Given an n-by-n matrix $A = (a_{ij})$ of non-negative integers, find the permutation $j_1, \ldots, j_n$ of the integers $1, \ldots, n$

[1]The work reported by this note was supported by the Office of Naval Research Logistics Project, Department of Mathematics, Princeton University.

*Manuscript received January 16, 1956.

421661 O - 57 - 4

that minimizes the sum $a_{1j_1} + \ldots + a_{nj_n}$. (If a maximum permutation sum is demanded, set $a = \max_{i,j} a_{ij}$ and replace A by $A' = (a'_{ij})$ with $a'_{ij} = a - a_{ij}$, for all i and j.) The Hungarian method is then based on two assertions:

A. The problem is unchanged if the matrix A is replaced by $A' = (a'_{ij})$, with $a'_{ij} = a_{ij} - u_i - v_j$ for constants u_i and v_j, and $i, j = 1, \ldots, n$.

B. (König's theorem) The minimum number of lines (rows and/or columns) needed to contain all of the zeros of A is equal to the maximum number of zeros that can be chosen, with no two on the same line.

As presented in [1], the algorithm divides into two routines. Routine I is essentially an iterative procedure for determining the minimum number of lines and the maximum number of zeros referred to in König's theorem; it is the possible variations in this routine that form the main subject of this paper. The input of the routine is a subset of the zeros of A, marked by asterisks, with no two asterisks in the same line. If the number of asterisks is n, then the desired minimum is zero, and the problem is solved by $j_1, \ldots, j_n$ where the asterisk in row i appears in column j_i, for $i = 1, \ldots, n$. If the problem is not solved (i.e., if there are $k < n$ asterisks) then the output of a single application of Routine I is one of the following disjoint alternatives:

Ia. A new set of $k + 1$ asterisks on zeros in A, with no two on the same line; or

Ib. A set C of $k < n$ lines which contain all of the zeros of A.

Routine I is repeated until there are n asterisks (and the problem is solved) or until there is an occurrence of Alternative Ib. In the latter event, let $h > 0$ be the minimum entry not appearing in the lines of C. Routine II calls for the addition of h to all entries in each line of C (if an entry appears in two lines, then 2h is added) then for the subtraction of h from every entry in A. Routine II leaves A non-negative but decreases the sum of all entries by an amount $nh(n - k)$. The only property of Ib needed for the finite termination of the combined algorithm is that the set C contains less than n lines. The original Routine I produces the exact minimum k; in the modification of Routine I proposed in the next section, the set C contains exactly $n - 1$ lines at every occurrence of Alternative Ib.

3. MODIFIED ROUTINE I

Search each column of A in turn for a 0^*. If a 0^* is found, proceed to the next column (if no columns remain then there are n asterisks and the problem is solved). If a 0^* is not found in the column, then the column is called pivotal and is searched for all of its 0's. If no 0 is found in the pivotal column, then the remaining $n - 1$ columns contain all of the zeros of A (Modified Alternative Ib). If there are 0's in the pivotal column, then their rows are searched in turn for a 0^*. If such a row contains no 0^*, then the 0 in that row and in the pivotal column is marked with an asterisk (Alternative Ia). If each 0 in the pivotal column has a 0^* in its row, then these rows are listed in any order: $i_1, \ldots, i_t$. We now construct a sequence of rows based on this initial segment. At step s of the construction, we search row i_s for a 0^*. If a 0^* is found, then the row indices of all 0's in its column are added to the sequence in any order and any indices already present are omitted. If no 0^* is found in row i_s, then a transfer (see [1] and Section 4 below) is possible. Namely, i_s was added to the sequence via a 0 in the column of a 0^* in an earlier row i_r of the sequence. If the asterisk is transferred from row i_r to row i_s in this column, then we have the same situation earlier in the sequence. Continuing, we ultimately free an asterisk in a row of the initial segment, $i_1, \ldots, i_t$. But then

an asterisk can be marked in this row of the pivotal column (Alternative Ia). In the remaining case, a finite sequence $i_1, \ldots, i_v$ is constructed with the following properties: Each row in the sequence contains a 0^* in some column. These columns, together with the pivotal column ($v + 1$ columns in all), contain 0's (marked or not) only in rows $i_1, \ldots, i_v$. Hence, the remaining $n - v - 1$ columns combined with the rows $i_1, \ldots, i_v$ constitute $n - 1$ lines that contain all zeros (Modified Alternative Ib).

In a rough attempt to evaluate the advantages of this variant, it may simplify the search for a transfer in Routine I at the possible expense of smaller changes in the sum of the entries in A on the next application of Routine II.

4. A GRAPH-THEORETICAL EQUIVALENT OF ROUTINE I

In this section, the problem solved by Routine I (and by the algorithm of Section 3) is given an independent statement and then rephrased as a problem in graph theory.

Let A be an n-by-n square array in which certain of the places are occupied by zeros. An assignment is a subset of k zeros, distinguished by asterisks, such that no two asterisks lie in the same line (row or column) of A. An assignment is complete if every zero lies in the line of some asterisk. A cover is a set of lines of A that contain all of the zeros in A; if a complete assignment is available, only the lines in which asterisks appear will be used in a cover. König's theorem asserts that the largest number of asterisks in an assignment is equal to the smallest number of lines in a cover. Such an assignment is called maximal; such a cover is called minimal. A transfer is possible when there is a sequence of asterisks at $(i_1, j_1), \ldots, (i_r, j_r)$ in A such that there are zeros at $(i_2, j_1), \ldots, (i_r, j_{r-1})$ and a zero at (i_1, j_0), for some j_0, with no asterisk in its column. The transfer removes the asterisks from $(i_1, j_1), \ldots, (i_r, j_r)$ and assigns them to $(i_1, j_0), \ldots, (i_r, j_{r-1})$. If there is a zero at (i_{r+1}, j_r), for some i_{r+1}, with no asterisk in its row, then the result of the transfer is an incomplete assignment (and another asterisk can be placed at (i_{r+1}, j_r)).

To each complete assignment for A, we shall associate an oriented graph G by the following rules:

The nodes V of G are in a 1-1 correspondence with the asterisks. A directed edge joins V_1 to V_2 if there is a zero in intersection of the column of the asterisk-node V_1 and the row of the asterisk-node V_2. The node V is distinguished as a source (sink) if there is a zero in the row (column) of V with no asterisk in its column (row). Note that a node can be both a source and a sink.

The equivalence of various concepts in the matrix A and the graph G is displayed in the table below:

Matrix A	Graph G
Transfer	Directed path (possibly void) originating at a source.
Transfer yielding incomplete assignment	Directed path (possibly void) from a source to a sink.
Cover	Assignment of + and/or - to each node so that every source (sink) is marked +(-), and no edge starts at a + and ends at a - .

The first two equivalences follow directly from the definitions. To prove the last equivalence, suppose that a cover is given for the matrix A in which the assignment is complete. Each asterisk is contained in a line of the cover; if it is covered by a row (column), mark the corresponding node by a + (-). An asterisk-node V is a source (sink) if, and only if, there is a zero in the row (column) of V without an asterisk in its column (row). This zero is covered; hence the node is marked with a + (-) if it is a source (sink). Each edge in G corresponds to an unmarked zero, and leads from an asterisk-node in its column to an asterisk-node in its row. For this zero to be covered, either the former is marked by a - or the latter by a + (Recall that when the available assignment is complete, only lines through asterisks appear in a cover.) To show the other half of the equivalence, consider the set of the rows of the asterisk-nodes marked with a + and of the columns of the asterisk-nodes marked with a -; under the conditions listed, it is asserted that this set is a cover. Clearly, only zeros without asterisks need be considered. If there is an asterisk in the row (column) of such a zero, but no asterisk in its column (row), then this asterisk-node is a source (sink) and is marked by a + (-). Hence the row (column) of the zero without an asterisk is in the set. If there is an asterisk in both the row and column of an unmarked zero, then there is an edge from the latter to the former. Since either the asterisk-node in the row is marked with a + or the asterisk-node in the column is marked with a -, the unmarked zero is covered.

THEOREM 1: Let A be a matrix with a complete assignment and let G be the associated oriented graph. Suppose each node of G is assigned + or - (and not both), so that every source (sink) is marked + (-) and no edge starts at a + and ends at a -. Then the corresponding cover is minimal.

PROOF: If there are k asterisks assigned in A, no cover can contain less than k lines. The cover corresponding to the given distribution of one + or - to each node of G contains exactly k lines. Q.E.D.

We now prove a purely graph-theoretic theorem that will form the basis of our discussion of Routine I (and, incidentally, prove König's theorem).

THEOREM 2: Let G be a directed graph in which two disjoint (and possibly void) sets of nodes have been distinguished as sources and sinks, respectively. Then exactly one of the following alternatives holds:

(a) There is a directed path in G from a source to a sink.

(b) Each node of G can be marked with one sign (+ or -), so that every source (sink) is marked + (-) and no edge starts at a + and ends at a -.

PROOF: Both alternatives cannot hold since, for any distribution of signs and any directed path from a source (+) to a sink (-), there must be an edge that starts at a + and ends at a -.

Let S be the set of nodes that are connected to some source by a directed path (starting at the source). If a sink lies in S then (a) holds. If a sink does not lie in S, then assign + to the nodes in S and - to the nodes not in S. This marks the sources and sinks as required by (b). Suppose that an edge starts at a node marked +, and hence in S. If the directed path from

a source to this node is extended by the edge in question, the terminus is connected to the source by the extended path. Hence the terminus is in S and is marked +. Q.E.D.

Remark 1: A proof of König's theorem now results as follows: It is a simple matter to construct a complete assignment of asterisks to the zeros in A (say, by assigning an asterisk to the first zero in each row which is not in the column of an asterisk that has been placed previously). If alternative (a) of Theorem 2 holds, a transfer yields an incomplete assignment and another asterisk can be placed. Repeating, alternative (a) can hold only a finite number of times (surely no more than n - k). When alternative (b) first occurs, we have a minimal cover with the same number of lines as there are assigned asterisks, and König's theorem is proved.

Remark 2: If there are no sources (sinks) then the theorem is trivially satisfied by marking all nodes - (+).

Remark 3: The construction of the proof can be altered in the following manner. Let S be as above. Let T be the set of nodes that are connected to a sink by a directed path (ending at the sink). If there is a node both in S and T, then (a) holds. Otherwise, mark the nodes in S (T) with + (-), and mark all nodes that are in neither S nor T with the same sign (either all + or all -).

5. VARIANTS OF ROUTINE I ON THE GRAPH G

The methods described below all amount to a construction of the set S or T in different manners. They will be defined in informal geometric terms, and the translation back into terms of the matrix A will be left for the reader.

1. (Kuhn [1]). In the search for a path from a source to a sink, start at any source and continue as far as possible along a directed path. If a sink is not reached, backtrack to the nearest branch point and try again. When the directed paths originating at one source are exhausted, try the next source. If all sources are exhausted and no sink has been reached, the set of all nodes that have been encountered is S.

2. (Ford-Fulkerson [2]). Start simultaneously from all sources and construct all directed paths with one edge (or less) originating at the sources. Extend these to all directed paths with two edges (or less). Continue fanning out from all sources until a sink is encountered or until S is exhausted.

3. (Hall [5]). Pick a column of A without an asterisk and call it pivotal. If there is no 0 in this column, then the remaining columns form a cover of n - 1 lines. Otherwise, each 0 is in the row of a source. Fan out from these sources (as in 2, above) until a sink is encountered or until the set of nodes connected to these sources by directed paths is exhausted. The cover then consists of the rows of the asterisk-nodes thus encountered completed by all columns except the pivotal column and the columns of asterisk-nodes already covered by rows (n - 1 lines in all).

4. (Munkres [6]). Mark all nodes both + and -. Then examine the nodes that have two signs, according to some preassigned order (say, from left to right in the matrix A). If such a node is a source (sink) or is connected from (to) a node marked only + (-) by a directed edge, remove the - (+) from the node and prime the edge used in the criterion (if any). If both signs are removed, a unique path from a source to a sink consisting of primed edges has been found. Otherwise, the examination is repeated until no more removals are indicated. At this point, the nodes marked + (-) constitute S (T) and the remaining double signs can be changed to -. (See Remark 3, above.)

Through the use of the graphical model, the reader can easily concoct variants of his own. Moreover, simple graphical examples show that none of the four methods given above is the best for all matrices A.

The method of Munkres sketched in Section 5 above is replaced in the revised version of his paper 6 by the following variant:

5. (Munkres 6). Mark all nodes - . Examine each source, sink, and edge in some pre-assigned order (since each of these objects is associated with one or more unassigned zeros in A, the natural order of these zeros can be followed). If a source marked - , or an edge connecting a + to a - , is found then change the - to a + (and prime the edge). If a sink marked + is found then a unique path consisting of primed edges from a source to a sink has been discovered. (We now note that, if the rows and columns associated with the current marks + and - are covered, each of the objects for which we are searching is associated with one or more uncovered zeros in A, and we can follow the natural order of these zeros.) Otherwise the process is repeated until all sources are marked + and no edge connects a + to a - (Alternative (b) of Theorem 2).

REFERENCES

[1] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," Naval Research Logistics Quarterly, Vol. 2, Nos. 1 and 2 (1955), pp. 83-97.

[2] L. R. Ford, Jr., and D. R. Fulkerson, "A Simple Algorithm for Finding Maximal Network Flows and an Application to the Hitchcock Problem," RAND Corporation, Santa Monica, P-743, 26 September 1955.

[3] M. M. Flood, "The Traveling-Salesman Problem," Operations Research, Vol. 4, (1956), pp. 61-75.

[4] T. S. Motzkin, "The Assignment Problem," Proc. of Symposia in Applied Math., Vol. VI - Numerical Analysis, pp. 109-125, McGraw-Hill, New York, 1956.

[5] Marshall Hall, Jr., "An Algorithm for Distinct Representatives," Amer. Math. Monthly, Vol. 63, (1956), pp. 716-717.

[6] J. Munkres, "Algorithms for the Assignment and Transportation Problems," to appear in the Journal of the Society for Industrial and Applied Mathematics (SIAM).

* * *

A FORMULATION OF THE DECISION PROBLEM FOR A CLASS OF SYSTEMS

Jack K. Hale
Remington Rand UNIVAC
St. Paul, Minn.
and
Ronald L. Reed
Gulf Research and Development Co.
Pittsburgh, Pa.

The fundamental problem of system analysis is to decide which is the best system of a given collection of systems. A precise formulation of this problem is made and a method of solution given which is, to a large extent, independent of the criterion employed.

1. INTRODUCTION

Suppose there are given several different systems, each of which serves the same general purpose. With respect to this purpose it is assumed that there can always be associated with each system a measure, V, of the value received when this system is employed, and a corresponding measure, C, of the cost. How does one decide which of these systems is, in some sense, superior to the others? This problem customarily has been resolved by the following procedure: A criterion D is selected, which is a function of V and C that appears to be particularly appropriate to the specific class of systems involved; and it is assumed that system i is superior to system j, providing $D(i) > D(j)$. For example, the specific criteria most frequently found in the literature are:

1. $D = V$ for (a) $C = C_0$, or for (b) $C \leq C_0$;
2. $D = -C$ for (a) $V = V_0$, or for (b) $V \geq V_0$;
3. $D = V - C$;
4. $D = V/C$.

All systems of the class are then compared under D, and that system having the largest value of D is called optimal (i.e., it is in this sense that one of the systems is considered superior to the others).

Certain aspects of the above approach appear unsatisfactory, namely: (1) to the authors' knowledge, no significant correspondence between classes of systems and criteria has been

*Manuscript received April 16, 1956.

established, (2) there is no possibility of arriving at an optimal solution composed of a mixture of the systems from the given class, and (3) it is frequently true that the numerical values of several of the parameters that enter into D are entirely unknown. There may also be other objections.

This paper attempts a characterization of the classes of systems for which maximization of D results in a function (preference function) that decides which systems are superior to the others independent of the specific manner in which D depends on V and C. Incidental to this characterization will be the result that for many systems the preference function obtained depends on substantially fewer parameters than appear in D.

To the best of the authors' knowledge, the first paper concerned with such a characterization is by R. W. Shephard (to be published). For the particular case of weapon systems appropriate to air defense (a generalization of example 6.1.2, below) and for the specific criteria 1(a), 2(a), and (4), Shephard has found a preference function by the use of methods different from those employed here.

We now summarize the results of the paper for the case in which a decision is to be made between two different systems and similar results are true for n different systems. This summarization is not mathematically precise, but it will aid in understanding the treatment which follows. We begin with an example.

Assume there are two types of weapons that may be used against a target of initial value V_0. Suppose that each strike with a weapon of type 1 (or type 2) is composed of a cell containing k_1 (or k_2) planes, only one of which carries a weapon. Also, suppose that p (or q) is the probability that a weapon of type 1 (or type 2) will not destroy the target and that c_1 (or c_2) is the cost of a weapon of type 1 (or type 2) plus the expected cost of delivery. These numbers will, in general, depend upon the cell size. The problem is to determine under what conditions one type of weapon is superior to the other. As pointed out previously, one usually compares these two systems in the following manner: carry out one strike with each weapon type, measure the value V and cost C associated with this strike, choose a criterion D(V, C), and if $D_1 > D_2$ then type 1 is superior to type 2. On the other hand, the procedure followed in this paper is the following: carry out x strikes using weapons of type 1 and y strikes using weapons of type 2, calculate the value V(x, y) and cost C(x, y) associated with these strikes, and choose any criterion D. By treating x and y as continuous variables (some of the implications of this assumption are given in 6.5), we use elementary calculus to determine the behavior of the surface D in terms of the functions V(x, y), C(x, y). In fact, if V_x, V_y, C_x, $C_y > 0$,* it is shown in (3.1) that along either of the curves, $V(x, y) = V_0$ or $C(x, y) = C_0$, the behavior of the surface D, for every function D with $D_V > 0$ and $D_C < 0$, is determined by the sign of the function $V_x/C_x - V_y/C_y$. If this function is positive, the larger value of D will occur in the direction of increasing x; if it is negative, the larger value of D will occur in the direction of increasing y. Consequently, one can use this function to determine the set of pairs (x, y) on which the maximum value of D must occur. If the pairs (x, y) are restricted to a particular set R, then the maximum value of D for every function D will occur either on the boundary of R or on the set of pairs (x, y) for which $V_x/C_x - V_y/C_y = 0$. In the following, the pair of functions $(V_x/C_x, V_y/C_y)$ is called a preference function.

*If $f = f(x_1, \ldots, x_n)$, then f_{x_i} is the partial derivative of f with respect to x_i.

Returning to the example, we have

$$V(x, y) = V_0(1 - p^x q^y)$$

$$C(x, y) = C_1 x + C_2 y$$

and $V_x/C_x = V_0 p^x q^y \, |\ln p|/C_1$, $V_y/C_y = V_0 p^x q^y \, |\ln q|/C_2$, and if $V_0, C_1, C_2 > 0$, then the sign of the function $V_x/C_x - V_y/X_y$ is determined by the function $|\ln p|/C_1 - |\ln q|/C_2$, which is independent of D, x, y and V_0. Consequently, if the only restriction in x and y is that they are positive, and if $|\ln p|/C_1 > |\ln q|/C_2$, then the maximum of D will occur for $y = 0$ and some x. In this case, we say that weapon type 1 is superior to weapon type 2. For this example, $V_x/C_x > V_y/C_y$ for every x, y, and, in such a case, the function $(V_x/C_x, V_y/C_y)$ is called an invariant preference function. General classes of value and cost functions which lead to invariant preference functions are discussed in Section 4; in particular, see Theorem 4.6.

2. FORMAL STATEMENT OF THE PROBLEM

For the purpose of this paper, a system $\mathfrak{S}$ is defined whenever a rule is established by means of which it is always possible to decide whether or not a given object belongs to the system. The collection of objects that make up a system fall into natural categories S_1, $S_2, \ldots$, according to their function or their physical identity, or both. This separation into categories is accomplished by using various rules, so that it follows that $S_1, S_2, \ldots$, are also systems; these are called subsystems. Each subsystem can frequently be separated in a similar fashion so that one arrives at a hierarchy of subsystems. The fundamental problem of system analysis is to select at every stage of a hierarchy that particular collection of subsystems which results in the best (optimal) system in some sense.

For simplicity, the discussion will be started by specializing to a two-stage hierarchy of subsystems. Let each subsystem S_i, $i = 1, 2, \ldots, N$, be characterized by the parameters $s_i = (s_{i_1}, \ldots, s_{i_n})$ so that $S_i = S_i(s_i)$, $i = 1, 2, \ldots, N$. Construct the system $\mathfrak{S} = \mathfrak{S}(S_1, \ldots, S_N, x_1, \ldots, x_N; \bar{r})$ from x_i identical subsystems of kind S_i, $i = 1, 2, \ldots, N$, where $\bar{r} = (r_1, \ldots, r_m)$ represents the collection of all parameters which affect $\mathfrak{S}$ but are exterior to all the subsystems S_i. Further, it is assumed that there is associated with every N-tuple $\bar{x} = (x_1, \ldots, x_N)$ a measure of the value $V(\bar{x}, \bar{s}, \bar{r})$ and a measure of the cost $C(\bar{x}, \bar{s}, \bar{r})$, where $\bar{s} = (s_1, \ldots, s_N)$. Finally, let $D(V, C)$ be an arbitrary function, increasing in V and decreasing in C. We shall introduce the following definitions:

2.1 Definition. Let $u_1, \ldots, u_n$ be n variables.

1. R^n is the region defined by $u_i \geq 0$, $i = 1, 2, \ldots, n$.
2. C^n is the class of all functions defined and continuous over R^n.
3. C_1^n is the class of all functions $H(u_1, \ldots, u_n)$ such that $H, H_{u_i} \in C^n$, $i = 1, 2, \ldots, n$.
4. C_2^n is the class of all functions $H(u_1, \ldots, u_n)$ such that $H \in C_1^n$ and $H_{u_i} > 0$, $i = 1, 2, \ldots, n$.

421661 O - 57 - 5

5. C_3^n is the class of all functions $H(u_1, \dots, u_n)$ such that $H \in C_2^n$ and the equation $H(0, \dots, 0, u_i, 0, \dots, 0) = H_0$ has a solution u_i^0, $i = 1, 2, \dots, n$ for every $H_0 \in [0, +\infty)$.

6. C^* is the class of all functions $H(u_1, u_2)$ such that $H \in C_1^2$, $H_{u_1} > 0$ and $H_{u_2} < 0$.

2.2 Definition. The system $\bar{x}_0$ is superior to the system $\bar{x}$ (in symbols $\bar{x}_0 > \bar{x}$) if[1] $D(\bar{x}_0) > D(\bar{x})$; $\bar{x}_0$ is not inferior to $\bar{x}$ (in symbols $\bar{x}_0 \geq \bar{x}$) if $D(\bar{x}_0) \geq D(\bar{x})$.

Let the functions $d_{S_i}(\bar{x}, \bar{s})$, $i = 1, 2, \dots, N$, be defined for a known set of parameters $\bar{s}$ and all $\bar{x} \in R^N$. Further, let a class $\{D\}$ of criteria be given, and suppose $R' \subset R^N$.

2.3 Definition. The N-tuple $\left(d_{S_1}, \dots, d_{S_N}\right)$ is a preference function with respect to R' and $\{D\}$, if for every $\bar{x} \in R'$ at which $d_{S_i} \neq d_{S_j}$ for at least one pair i, j, there exists a point $\bar{x}_1 \in R'$ such that $\bar{x}_1 > \bar{x}$ for every $D \in \{D\}$.

2.4 Definition. The N-tuple $d_{S_1}, \dots, d_{S_N}$ is an invariant preference function with respect to R' and $\{D\}$ if it is a preference function for which there exists a permutation of $(1, 2, \dots, N)$ such that for all $\bar{x} \in R'$, $d_{S_1} > d_{S_2} > \dots > d_{S_N}$.

Within the framework of the preceding discussion, the theorems which follow characterize a class of systems $\{S\}$ for which the preference function (or invariant preference function) is independent of the specific functional dependence of D on V and C. It should be noted that V and C need not be measures of value and cost in the following. The discussion is merely a treatment of the behavior of a certain class of surfaces $\{D[V(\bar{x}), C(\bar{x})]\}$.

3. PREFERENCE FUNCTIONS

With respect to the theorems to be proved, the components of $\bar{x}$ will always be treated as continuous variables. By means of an example (Section 6.5), some of the difficulties encountered if the components of $\bar{x}$ are considered integral will be illustrated.

3.1 Theorem

If[2] $V, C \in C_3^n$, then the N-tuple $\left(V_{x_1}/C_{x_1}, V_{x_2}/C_{x_2}, \dots, V_{x_N}/C_{x_N}\right)$ is a preference function with respect to R' and $\{D | D \in C^*\}$, where R' is either of the following:

1. Any subset R' of any member Γ of the family of surfaces $\{V(\bar{x}) = V_0\} \cup \{C(\bar{x}) = C_0\}$, where R' is open in $\Gamma \cap$ (Interior of R^N).

---

[1] In the following, the symbol $D(\bar{x})$ represents $D(V(\bar{x}, \bar{s}, \bar{r}), C(\bar{x}, \bar{s}, \bar{r}))$, and where convenient the dependence of V and C on $\bar{s}$ and $\bar{r}$ is omitted. Further, the distinction between the parameters $\bar{s}$ and $\bar{r}$ is not essential, and they will all be considered as represented by the single symbol $\bar{s}$.

[2] $a \in B$ if the element a belongs to the set B. $a \in A \cup B$ if either $a \in A$ or $a \in B$. $a \in A \cap B$ if $a \in A$ and $a \in B$. $A \subset B$ if each element of the set A belongs to the set B.

2. Any set R' which can be represented as a union of sets defined in 1, and in particular any open subset of the interior of R^N.

PROOF: Let $V, C \in C_3^N$, and suppose Γ is one of the surfaces $\{C(\bar{x}) = C_0\}$ (for $\Gamma \in \{V(\bar{x}) = V_0\}$ the proof is the same). It is sufficient to let $\bar{x}_0$ be any point of $R' \subset \Gamma$ for which $V_{x_i}/C_{x_i} > V_{x_j}/C_{x_j}$ for at least one pair i, j. Since $C \in C_2^N$, the function $C_{x_j} \neq 0$ at the point $\bar{x}_0$. It then follows from the implicit function theorem that the relation $C(\bar{x}) - C_0 = 0$ uniquely defines x_j as a continuous function of $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_N$ throughout a certain region in R^N about $\bar{x}_0$ and, consequently, throughout a set R'' containing $\bar{x}_0$ and open in R'. Furthermore, $\partial x_j/\partial x_i = -C_{x_i}/C_{x_j}$ everywhere in R''. Hence, for every $D \in C^*$,

$$D_{x_i} = D_V\left[V_{x_i} + V_{x_j}\left(-C_{x_i}/C_{x_j}\right)\right] = D_V C_{x_i}\left[V_{x_i}/C_{x_i} - V_{x_j}/C_{x_j}\right]$$

everywhere in R'', and $D_{x_i}(\bar{x}_0) > 0$. Since D is continuous in $\bar{x}$ throughout R^N and R'' is open in R', a point $\bar{x}_1 \in R''$ exists such that $\bar{x}_1 > \bar{x}$ for every $D \in C^*$. Since $\bar{x}_0$ is an arbitrary point of R', $\left(V_{x_1}/C_{x_1}, \ldots, V_{x_N}/C_{x_N}\right)$ is a preference function with respect to $\{D|D \in C^*\}$ and R' as defined in hypothesis 1.

If R' is a set as defined in hypothesis 2, the theorem is clearly true, because through any point in R' there passes a surface of the family $\{C(\bar{x}) = C_0\}$.

3.2 Corollary

If R' is a subset of the $2 \leq m < N$ dimensional set $\left(x_{i_j} > 0, j = 1, 2, \ldots, m\right)$, then $\left(V_{x_{i_1}}/C_{x_{i_1}}, \ldots, V_{x_{i_m}}/C_{x_{i_m}}\right)$ is a preference function with respect to R' and $\{D|D \in C^*\}$, where $\left(x_{i_1}, \ldots, x_{i_m}\right) \in (x_1, \ldots, x_N)$.

3.3 By inspecting the proof of Theorem 3.1, it is clear that it need only be required that $D_V > 0$, $D_C \leq 0$ [or $D_C < 0$, $D_V \geq 0$], provided that Γ is restricted to the family of surfaces $\{C(\bar{x}) = C_0\}$ [or $\{V(\bar{x}) = V_0\}$]. This permits the discussion of functions D which are independent of C [or V].

3.4 Further, Γ need not have been restricted to curves of the form $V = V_0$ or $C = C_0$. It is easily shown that the result is the same when Γ is any curve having a slope always lying between the slopes of these constant value and cost curves. In fact, this is the largest class of curves Γ for which the theorem is true. Since the additional variety in the sets R' which can be admitted is not greatly increased, use of only the constant-value and constant-cost curves will be continued.

The preference function obtained in the preceding theorem can be used to reduce the possible sets on which the maximum of D can occur. In the following theorem (3.5), R' is any set satisfying property 1 or 2 of Theorem 3.1, i.e., $R' = U_\Gamma (R' \cap \Gamma)$ where Γ ranges over an appropriate subset of $\{V = V_0\} \cup \{C = C_0\}$. The discussion is also restricted to two variables x and y, and it is assumed that every curve Γ is intersected by the curves $V_x/C_x - V_y/C_y = 0$ at most a finite number of times.

3.5 Theorem

For every $D(V, C) \in C^*$ with $V, C \in C_3^2$, every maximum[3] of D on $Cl(R')$ occurs at a point (x, y), which lies on one of the curves $V_x/C_x = V_y/C_y$ or on the set[4] $Cl\{U_\Gamma [Cl(\Gamma \cap R') - (\Gamma \cap R')]\}$.

PROOF: Since the curves $B(x, y) \equiv V_x/C_x - V_y/C_y = 0$ intersect Γ in at most a finite number of points, it follows that these curves divide the set $R' \cap \Gamma$ into a finite number of sets γ_α, such that each γ_α is open in Γ and the function B has fixed sign (but may vanish) on γ_α.

Continuity of the functions B and D on Γ implies their continuity on $Cl(\gamma_\alpha)$ for each α, and since B has fixed sign on γ_α, it has fixed sign on $Cl(\gamma_\alpha)$. Consequently, the maximum of D on $Cl(\gamma_\alpha)$ must lie on the set $Cl(\gamma_\alpha) - \gamma_\alpha$, which consists of the endpoints of $Cl(\gamma_\alpha)$. Since the set of α's is finite, it follows that the maximum of D on $Cl(R' \cap \Gamma)$ must lie on the set $U_\alpha [Cl(\gamma_\alpha) - \gamma_\alpha] \equiv S_\Gamma$. It is clear that these points will be either zeros of B or boundary points of R'.

Furthermore, $R' = U_\Gamma (\Gamma \cap R')$ implies that for any $(x, y) \in R'$ there is an (x_0, y_0), depending on (x, y) and contained in $U_\Gamma S_\Gamma$, such that $D(x_0, y_0) \geq D(x, y)$. But this implies that the absolute maximum of D on $Cl(R')$ must belong to the closed set $Cl\{U_\Gamma S_\Gamma\}$. Therefore, since each set S_Γ can be written $S_\Gamma = Cl(\Gamma \cap R') - \Gamma \cap R' + \Gamma \cup (B = 0)$, it follows that $Cl\{U_\Gamma S_\Gamma\} = Cl\{U_\Gamma [Cl(\Gamma \cap R') - \Gamma \cap R']\} + Cl\{U_\Gamma [\Gamma \cap (B = 0)]\}$. The observation that $Cl\{U_\Gamma [\Gamma \cap (B = 0)]\}$ lies on the curves $V_x/C_x - V_y/C_y = 0$ concludes the proof.

3.6 Theorem 3.5 can be sharpened somewhat. It is clear from the proof that the maximum of D can occur only on certain of the curves $B(x, y) = 0$, but not on all of them. It is only necessary to consider those portions of these curves where B changes sign along Γ curves at the intersection points. Even this class is too large and can be reduced further by elimination of the places where a change in sign of B implies a minimum rather than a maximum.

3.7 To illustrate more clearly the implications of the above statements, suppose the curve $V_x/C_x = V_y/C_y$ intersects each of the curves $C(x, y) = C_0$ in, at most, one point, and suppose these curves are similar to those shown in Figures 1a and 1b. If $V_x/C_x < (>) V_y/C_y$ for $x < (>) x_1$, then the maximum of D in R^2 must occur on the hatched lines in Figure 1a. If $V_x/C_x > (<) V_y/C_y$ for $x < (>) x_1$, then the maximum of D in R^2 must occur on the hatched lines in Figure 1b.

3.8 Theorem 3.5 may be generalized to the case of N variables. In such a case, the conclusion is that the maximum of D occurs for some $\bar{x}$ which belongs either to the set $Cl\{U_\Gamma [Cl(\Gamma \cap R') - (\Gamma \cap R')]\}$ or to one of the sets of points where either $d_{S_i} = d_{S_j}$, $d_{S_i} = d_{S_j} = d_{S_k}$, or, ... , or $d_{S_1} = d_{S_2} = \ldots = d_{S_N}$.

[3] In unusual situations, there may be more than one maximum.

[4] By Cl(A) is meant the closure of A in R^2. In most cases, the set mentioned in the theorem will consist merely of those points which are added to the boundary of R' by closing R'.

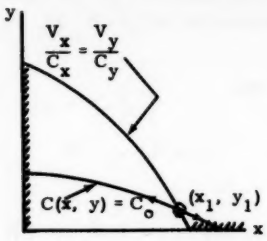

Figure 1a

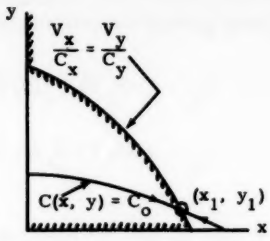

Figure 1b

4. INVARIANT PREFERENCE FUNCTIONS

In this section, the invariant preference functions (see Section 2.4) for a two-fold hierarchy will be treated, and for simplicity the discussion will commence with V and C defined only over the two-dimensional plane of variables x, y.

4.1 Even though, in general, the preference function depends on (x, y), it may happen that $V_x/C_x - V_y/C_y$ does not change sign for any (x, y) in R^2. In such a case, it follows from Theorem 3.1 that the pair $(V_x/C_x, V_y/C_y)$ is an invariant preference function, and it is clear from Theorem 3.5 that the maximum of D lies on the y-axis or on the x-axis, according as $V_x/C_x \leq V_y/C_y$ or $V_x/C_x \geq V_y/C_y$, respectively, for all (x, y) in R. Thus, with no side conditions on x and y, an invariant preference function leads to a preference, independent of the functional dependence of D upon V and C, of a pure system over all other systems. This is an important class of systems and therefore warrants further study.

If $V, C \in C_3^2$, the pair $(V_x/C_x, V_y/C_y)$ is an invariant preference function if, and only if, the expression $V_xC_y - V_yC_x = f(x, y) \neq 0$ for all (x, y). To describe a class of functions V, C which lead to invariant preference functions, it would be sufficient to solve this underdetermined partial differential equation for some $f(x, y) \neq 0$. This means C could be considered an arbitrary function of x, y, and $V = V(C, x, y)$ determined in such a way that the above equation is satisfied. Classes of solutions can easily be obtained in this manner, but given two functions V and C, it is usually easier to show $f(x, y) \neq 0$ throughout R^2 than to verify the required functional dependence of V upon C, x, and y.

However, some classes for which it is easy to verify this functional dependence may be obtained by finding those functions V, C which satisfy the relationship $V_x/C_x = (b/a)(V_y/C_y)$, i.e., $aV_xC_y - bV_yC_x = 0$. This differential equation is equivalent to the statement that the Jacobian of the functions V(x, y) and C(bx, ay) vanishes throughout R^2. Consequently, for every given C, a function W exists such that V and C are functionally related through $W[V(x, y), C(bx, ay)] = 0$. Conversely, it is clear that $V = U[C(bx, ay)]$ satisfies this differential equation for arbitrary functions U. There are two subclasses of these solutions which seem to be of interest, namely, the solutions which correspond to functions V and C which satisfy either (1) $V_x/V_y = b$, $C_x/C_y = a$, or (2) $V_x/C_x = b$, $V_y/C_y = a$. It is easy to show that $V = W(bx + y)$, $C = U(ax + y)$, with W and U arbitrary, are the general solutions of (1); and $V = bg(x) + ah(y)$, $C = g(x) + h(y)$, with g and h arbitrary, are the general solutions of (2). In the remainder of this section, (1) will be generalized to the case of N variables.

4.2 Definition. The functions V and C belong to the class P^N on a set $R' \subseteq R^N$ if $V, C \in C_3^N$ and, whenever their partial derivatives are written for all $\bar{x} \in R'$ as

$$V_{x_i} = f_i(\bar{s}, \bar{r})\, F_i(\bar{x}, \bar{s}, \bar{r}) \text{ and}$$

$$C_{x_i} = g_i(\bar{s}, \bar{r})\, G_i(\bar{x}, \bar{s}, \bar{r}), \qquad i = 1, 2, \ldots, N,$$

where the functions f_i and g_i are all positive, the integers $i = 1, 2, \ldots, N$ can be partitioned into $1 \leq k \leq N$ distinct nonempty subsets I_n such that for $n = 1, 2, \ldots, k$:

1. $F_i = F_j$ and $G_i = G_j$ for all $(i, j) \in I_n$ and
2. $f_{i_0(n)}/g_{i_0(n)} > f_i/g_i$ for some $i_0(n) \in I_n$ and all $i \neq i_0(n) \in I_n$.

The proof of the main theorem on functions $V, C \in P^N$ requires the following two lemmas, the first of which may be found elsewhere.[5]

4.3 Lemma

A necessary and sufficient condition that n functions of n + p independent variables be connected by a relation which does not involve these variables is that every one of the Jacobians of these n functions, with respect to any n of the independent variables, should vanish identically.

For the following lemma it is convenient to introduce the notation $\phi_n = \sum_{i \in I_n} f_i x_i$, $\psi_n = \sum_{i \in I_n} g_i x_i$, $n = 1, 2, \ldots, k$ and $\phi = (\phi_1, \ldots, \phi_k), \psi = (\psi_1, \ldots, \psi_k)$, where I_n is defined in Definition 4.2.

4.4 Lemma

Suppose $V(x_1, \ldots, x_N) \in C_1^N$ and each ϕ_n, $n = 1, 2, \ldots, k$ contains at least two of the variables $x_1, \ldots, x_N$. Then the Jacobian of the $k + 1$ functions $\phi_1, \ldots, \phi_k, V$ with respect to any $k + 1$ independent variables $x_{i_1}, \ldots, x_{i_{k+1}}$ is either identically zero or of the form $A\left(f_j V_{x_i} - f_i V_{x_j}\right)$, where A is a constant and the indices $i, j \in (i_1, \ldots, i_{k+1}) \cap I_n$ for some n. A similar result holds for the functions $\psi_1, \ldots, \psi_k$ and some function $C \in C_1^N$.

PROOF: It is clear that $\partial\phi_n/\partial x_{i_k} = 0$ if $i_k \notin I_n$, and $\partial\phi_n/\partial x_{i_k} = f_{i_k}$ if $i_k \in I_n$. Suppose there is a permutation of the indices $i_1, \ldots, i_{k+1}$ such that $\left(i_1, \ldots, i_{\alpha_1}\right) \in I_1$, $\left(i_{\alpha_1 + 1}, \ldots, i_{\alpha_2}\right) \in I_2, \ldots, \left(i_{\alpha_{k-1} + 1}, \ldots, i_{k+1}\right) \in I_k$. If some I_n contains none of the indices i_j, then

[5] Goursat, E. J. B., A Course in Mathematical Analysis, New York, Ginn and Co., 1904, Vol. I, p. 56.

choose $\alpha_n = \alpha_{n-1}$ and omit I_n from the above sequence. Furthermore, abbreviate i_j by j, $j = 1, 2, \ldots, k+1$. Then

$$\frac{\partial(\phi_1, \ldots, \phi_k, V)}{\partial(x_1, \ldots, x_{k+1})} = \begin{vmatrix} f_1, \ldots, f_{\alpha_1} & & & & \\ & f_{\alpha_1+1}, \ldots, f_{\alpha_2}, & & & \\ & & \ddots & & \\ & & & f_{\alpha_{k-1}+1}, \ldots, f_{k+1} & \\ V_{x_1}, \ldots, V_{x_{\alpha_1}}, & V_{x_{\alpha_1+1}}, \ldots, V_{x_{\alpha_2}}, & \ldots, & V_{x_{\alpha_{k-1}+1}}, \ldots, & V_{x_{k+1}} \end{vmatrix},$$

where all blank spaces in this determinant represent zeros. Expand this determinant along the last row and consider the cofactor of V_{x_1}. This cofactor,

$$(-1)^{k+1} \begin{vmatrix} f_2, \ldots, f_{\alpha_1}, & & & \\ & f_{\alpha_1+1}, \ldots, f_{\alpha_2}, & & \\ & & \ddots & \\ & & & f_{\alpha_{k-1}+1}, \ldots, f_{k+1} \end{vmatrix},$$

is zero except when $\alpha_1 = 2$, $\alpha_2 = 3, \ldots, \alpha_{k-1} = k$. If these conditions are satisfied, then

$$\frac{\partial(\phi_1, \ldots, \phi_k, V)}{\partial(x_1, \ldots, x_{k+1})} = \begin{vmatrix} f_1, & f_2, & & & \\ & & f_3, & & \\ & & & \ddots & \\ & & & & f_{k+1} \\ V_{x_1}, & V_{x_2}, & V_{x_3}, & \ldots, & V_{x_{k+1}} \end{vmatrix} = (-1)^{k+1} \left(V_{x_1} f_1 - V_{x_2} f_2\right) f_3 \cdots f_{k+1}.$$

This discussion for V_{x_1} holds true for any V_{x_i}, and the desired result is obtained. The proof is similar for $\psi_1, \ldots, \psi_k$ and C.

4.5 Theorem

If $V, C \in P^N$ on a set $R' \subset R^N$ (see 4.2), then there exist functions U, W such that $U(V, \phi) = 0$ and $W(C, \psi) = 0$ where the functions ϕ, ψ are the same as those in lemma 4.4. Conversely, if $V = U(\phi)$ and $C = W(\psi)$ where $U, W \in C_3^1$ but are otherwise arbitrary, then $V, C \in P^N$ on R^N.

PROOF: Let every set I_n mentioned in Definition 4.2 contain at least two variables. If the functions $V, C \in P^N$ on a set $R' \subset R^N$, then they must satisfy the system of equations

$$\left.\begin{aligned} f_j V_{x_i} - f_i V_{x_j} &= 0 \\ g_j C_{x_i} - g_i C_{x_j} &= 0 \end{aligned}\right\} \quad i, j \in I_n,\ n = 1, 2, \ldots, k .$$

Conversely, if $V, C \in C_3^N$ and these equations are satisfied, then $V, C \in P^N$; for, if $V_{x_i} = f_i \xi(\bar{x})$, $V_{x_j} = f_j \eta(\bar{x})$, $f_i, f_j > 0$, then $f_j V_{x_i} - f_i V_{x_j} = f_i f_j (\xi - \eta) \neq 0$ unless $V \in P^N$. Similarly, $C \in P^N$.

The remainder of the proof for the case where each I_n contains at least two variables follows directly from Lemmas 4.3 and 4.4. If some of the sets I_n contain only one variable, the theorem is proved in the same manner by applying Lemmas 4.3 and 4.4 only to those sets I_n which contain at least two variables.

If Definition 4.2 is combined with Corollary 3.2, the following theorem is immediate.

4.6 Theorem

If $V, C \in P^N$ on R^N and $D \in C^*$, then the tuple $(\{f_i/g_i\},\ i \in I_n)$ is an invariant preference function over any set $R' \subseteq \{\bar{x} \in R^N \mid x_i$ fixed for all $i \notin I_n\}$ which satisfies the hypotheses 1 and 2 of Theorem 3.1.

4.7 If $V, C \in P^N$ (see 4.2) on R^N, and $D \in C^*$, then there is a system $\bar{x}^* \in R^N$ having all components zero, except $x_{i_o(n)}$ [with $i_o(n)$ as defined in 4.2], $n = 1, 2, \ldots, k$, such that $\bar{x}^* > \bar{x}$ for all $\bar{x} \in R^N$.

In this manner the N dimensional problem is reduced to a $k < N$ dimensional problem, and the preference function in this k dimensional space is $\left(V_{x_{i_o(1)}}/C_{x_{i_o(1)}}, \ldots, V_{x_{i_o(n)}}/C_{x_{i_o(n)}}\right)$.

4.8 If $n = 1$, then the n-tuple $(f_1/g_1, \ldots, f_N/g_N)$ is an invariant preference function over R^N. If, in addition, $f_i(\bar{s}, \bar{r}) = f(s_i, \bar{r})$, $g_i(\bar{s}, \bar{r}) = g(s_i, \bar{r})$, $i = 1, 2, \ldots, N$, then the subsystem $S_i(s_i)$ which maximizes the ratio $f(s_i, \bar{r})/g(s_i, \bar{r})$ with respect to s_i is not inferior to any other subsystem. Consequently, the system design may be based on this ratio.

5. SEVERAL STAGES

The foregoing discussion covered exclusively the problem of a two-stage hierarchy. This section will indicate how a three-stage hierarchy can be treated by a brief glance over the case where a particular invariant preference function exists in the second stage of the system.

Suppose (see Section 4.8) that $\left(d_{S_1}, d_{S_2}\right) = \left(f_{S_1}/g_{S_1}, f_{S_2}/g_{S_2}\right)$, where $f_{S_i} = f(\bar{s}_i; \bar{r})$ and $g_{S_i} = g(\bar{s}_i; \bar{r})$, $i = 1, 2$. Since $d_{S_1} > d_{S_2}$ implies S_1 superior to S_2 (in the sense that an x_o exists such that $D(x_o, 0) > D(x_1, x_2)$ for all $(x_1, x_2) \in R^2$), maximization of the ratio f_{S_1}/g_{S_1} will determine an optimal system for the two-stage hierarchy. Hence, to determine the optimal

combination of subsystems for a three-stage hierarchy, it is sufficient to determine the combination of subsystems of S_1 that will maximize f_{S_1}/g_{S_1} and then determine the number of subsystems S_1 that will maximize D. Suppose then that $S_1 = S_1(s_1)$ consists of y_1 identical subsystems of kind $T_1 = T_1(t_1)$ and of y_2 identical subsystems of kind $T_2 = T_2(t_2)$. Then $\bar{s}_1 = \bar{s}_1(y_1, y_2, t_1, t_2, \bar{m})$, where $\bar{m}$ is the collection of parameters that affect S_1 but are exterior to T_1 and T_2, and the functions f_{S_1} and g_{S_1} depend upon the variables y_1, y_2 and the fixed parameters $t_1, t_2, \bar{m}$, and $\bar{r}$. The problem now is to maximize the function $D_{S_1} = f_{S_1}(y_1, y_2)/g_{S_1}(y_1, y_2)$ with respect to y_1 and y_2. This situation is exactly analogous to the preceding problem with f_{S_1} and g_{S_1} taking the roles of V and C, respectively; thus, if it is assumed that $f_{S_1}, g_{S_1} \in C_3^2$, the same kind of results can be expected.

It should be noticed that the actual quantities of the optimal subsystems can be determined for all stages of the hierarchy except the second, since the complete criteria for these stages are precisely known and are independent of D.

6. EXAMPLES

6.1 $V = V(p^x q^y)$, $C = C(C_1 x + C_2 y)$.

If $V = V(p^x q^y)$, $V' < 0$, $0 < p, q < 1$, $C = C(C_1 x + C_2 y)$, $C_1, C_2 > 0$, $C' > 0$, then V_x, V_y, $C_x, C_y > 0$, and, furthermore, $V = V(e^{x \ln p + y \ln q}) = U(x|\ln p| + y|\ln q|)$. Therefore, it follows, from Section 4.8, that the pair

$$(1) \qquad (|\ln p|/C_1, \ |\ln q|/C_2)$$

is an invariant preference function with respect to R^2 and $\{D|D \in C^*\}$. It should be emphasized that to obtain the preference function it is sufficient to know the form of the variables on which V and C depend (e.g., it is sufficient to know that C depends on x and y only through the form $C_1 x + C_2 y$; it is not necessary to know that C equals $C_1 x + C_2 y$), and knowledge of the actual functions is entirely unnecessary. Further, it is interesting to notice that the preference depends only on p, q, C_1, and C_2, independent of all other parameters which may enter into the expressions for V and C. In the examples which follow, the complete functions V and C can easily be written. In very complicated cases, however, this is not always practical, even though it is sometimes easy to reason merely that V and C are functions of $p^x q^y$ and $C_1 x + C_2 y$, respectively.

6.1.1 Offensive Bombing Campaign

This example was discussed in the introduction. We remark here only that it is now a special case of 6.1. It should be noticed that the system which is superior to all other systems could be obtained by maximizing the ratio $|\ln p|/c_1$ with respect to the parameters determining p and c_1. For example, if p and c_1 depend only upon k_1 and the characteristic parameters of the weapon, one could determine the optimal cell size for a given weapon by maximizing the ratio $|\ln p|/c_1$ with respect to k_1.

421661 O - 57 - 6

6.1.2 Air Defense

Assume that the defense of an installation of initial value v_0 is carried out by employing x weapons of type 1 and y weapons of type 2 against each attacking aircraft in such a way that each weapon affects only the aircraft for which it is intended. Assume further that the attack is made in K strikes, each composed of a cell containing k aircraft, only one of which carries a weapon. The problem again is to determine the circumstances under which one type of weapon is not inferior to the other. The value V expected to be received from the expenditure of x weapons of type 1 and y weapons of type 2 is defined to be the value of the installation expected to be saved, plus the value of the planes expected to be destroyed. If p (or q) is the probability that a weapon of type 1 (or type 2) will not destroy an enemy aircraft, and if C_1 (or C_2) is the expected cost of a weapon of type 1 (or type 2), then $V = V(p^x q^y)$ and $C = C(C_1 x + C_2 y)$.

Therefore, the preference function is the same as in the preceding example, but the expressions for p, q, C_1, and C_2 will be different.

6.2 $V = V(p^x q^y)$, $C = C(\mu^x \nu^y)$

If $V = V(p^x q^y)$, $C = C(\mu^x \nu^y)$, $V' < 0$, $C' < 0$, $0 < p, q, \mu, \nu < 1$, then $V_x, V_y, C_x, C_y > 0$ and, furthermore, $V = U(x|\ln p| + y|\ln q|)$, $C = W(x|\ln \mu| + y|\ln \nu|)$. Therefore, it follows from Section 4.8 that the pair $(\ln p/\ln \mu, \ln q/\ln \nu)$ is an invariant preference function with respect to R^2 and $\{D|D \in C^*\}$.

6.2.1 Antisubmarine Campaign

An aircraft carrier is used as a home base for dispatching cells of $x + y + 1$ planes in search of enemy submarines. One of these planes carries search equipment, x of the planes each carry a single weapon of type 1, and y of the planes each carry a single weapon of type 2. Each cell goes its own way and searches for submarines independently of the other cells. The problem is to determine under what conditions one type of weapon is superior to the other.

The object is to destroy submarines for the purpose of saving the shipping and shore installations which the submarine would attack if it survived. In carrying out this program there is involved a certain cost, namely, the carrier may be lost because of accidents of planes carrying weapons on take-off or landing, and the carrier and land storage base are both subject to destruction from predetonation of weapons in storage. It is now desired to determine the specific manner in which V and C depend upon x and y.

The value received from destroying submarines is the difference between the value $V(x, y)$ of ships, shore installations, etc. (all targets which submarines can attack) expected to remain if the planes attack submarines with x and y weapons on each mission and the value $V(0, 0)$ expected to remain if no weapons are employed against submarines (this is independent of x and y). The expression $V(x, y)$ is determined by the number of submarines, the number of times N any particular submarine is encountered during a cruise, the probability of survival $A(x, y)$ during each encounter, and the value V_s the submarine is expected to destroy between encounters. The function $V(x, y)$ is therefore a function of $A(x, y)$, since N depends (in a complicated manner) upon x, y only through the function $A(x, y)$, and V_s is independent of x, y. Furthermore, the probability of survival $A(x, y)$ during each encounter is given by $A(x, y) = p^x q^y$ where p (or q) is the probability that a weapon of type 1 (or type 2) does not destroy the submarine. Finally, it is seen that $V = V(x, y) - V(0, 0) = U(p^x q^y) = W(x|\ln p| + y|\ln q|)$, where W is (in principle) a known function.

Now consider the cost C. The expected carrier cost is the value of the carrier and its contents, multiplied by 1 minus the probability, P_c, that the carrier survives. If P_d is the probability that the weapons do not predetonate in storage and $P_t(x, y)$, $P_e(x, y)$ are the probabilities of successful take-off and landing, respectively, then $P_c = P_d P_t^m P_e^n$, where m is the number of take-offs and n is the number of landings. Assume that $n = m\delta$, where δ is the probability of not detecting a submarine on a mission. Further, P_d is assumed independent of x and y, on the basis that in either of the storage areas the total number of weapons does not vary much during a cruise and the variation in predetonation probability from weapon to weapon is small. If, in addition, μ_1 (or ν_1) is the probability that a weapon of type 1 (or type 2) is loaded on the plane and takes off without a detonation, μ_2 (or ν_2) is the probability that a plane carrying a weapon of type 1 (or type 2) lands and is unloaded without a detonation; then $P_t = \mu_1^x \nu_1^y$, $P_e = \mu_2^x \nu_2^y$, and the expected cost of the carrier is a function of $(\mu_1 \mu_2^{\delta})^x (\nu_1 \nu_2^{\delta})^y$. If the expected cost of the land storage base is assumed constant, and the cost of planes lost by predetonation while on a mission is ignored in comparison to carrier cost, then the total expected cost $C = C(x|\ln \mu_1 \mu_2^{\delta}| + y|\ln \nu_1 \nu_2^{\delta}|)$, where C is (in principle) a known function. Finally, from Section 6.2, the invariant preference function for this problem is given by $[\ln p/\ln (\mu_1 \mu_2^{\delta}), \ln q/\ln (\nu_1 \nu_2^{\delta})]$.

In the preceding example, it is of particular interest to observe that the weapon choice is independent of those parameters which would be almost impossible to evaluate. For example, the cost of carrier and contents (including lives) and the value of the destruction a submarine could accomplish if it survived do not appear in the preference function.

6.3 $$V = V\left(\sum_{k=1}^{N} a_k x_k - \prod_{k=1}^{N} p_k^{x_k}\right), \quad C = C\left(\sum_{k=1}^{N} C_k x_k\right).$$

If $V' > 0$, $C' > 0$, $a_i, C_i > 0$, $0 < p_i < 1$, $i = 1, 2, \ldots, N$, then $V_{x_i}, C_{x_i} > 0$, $i = 1, 2, \ldots, N$. Therefore, it follows from Section 3.1 that the N-tuple

(2) $$\left(d_{S_1}, d_{S_2}, \ldots, d_{S_N}\right) = \left[\left(a_k + |\ln p_k| \prod_{j=1}^{N} p_j^{x_j}\right) \Big/ C_k;\ k = 1, 2, \ldots, N\right]$$

is a preference function with respect to $\{D|D \in C^*\}$ and any set R' satisfying the conditions of Theorem 3.1. Furthermore, when $R' = \text{Interior } (R^N)$ one can conclude (see Section 3.8) that the maximum of D must occur for some $\bar{x}$ which belongs either to the boundary of R^N or to the set S of all points where either $d_{S_i} = d_{S_j}$, or $d_{S_i} = d_{S_j} = d_{S_k}$, or, ..., or $d_{S_1} = d_{S_2} = \ldots = d_{S_N}$.

The relation $d_{S_i} = d_{S_j}$ implies

$$\left[a_i + |\ln p_i| \prod_{k=1}^{N} p_k^{x_k}\right] \Big/ C_i = \left[a_j + |\ln p_j| \prod_{k=1}^{N} p_k^{x_k}\right] \Big/ C_j,$$

or

(3) $$\sum_{k=1}^{N} x_k |\ln p_k| = -\ln \left[(a_j/C_j - a_i/C_i)/(|\ln p_i|/C_i - |\ln p_j|/C_j)\right].$$

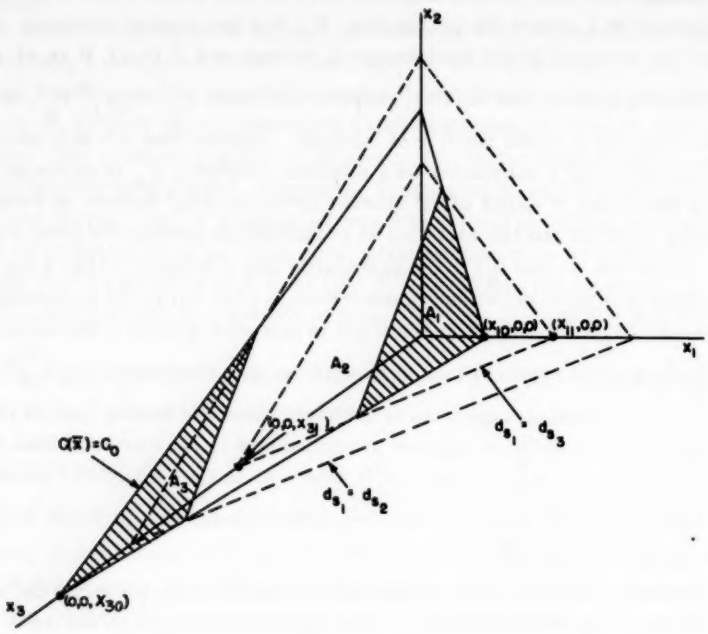

Figure 2a

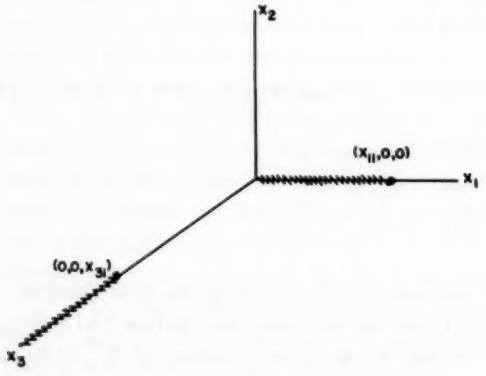

Figure 2b

According to Eq. 3, the conditions $d_{S_i} = d_{S_j}$ $(i, j = 1, 2, \ldots N, i \neq j)$ define a family of parallel hyperplanes. From this it follows that $d_{S_i} = d_{S_j} = d_{S_k}$, $i \neq j \neq k$ is never satisfied, so that the maximum of D must occur either on one of the planes defined by Eq. 3 or on the boundary of R^N.

Specific values of the parameters entering in this problem will in general allow one to restrict the maximum of D to a set smaller than that mentioned above. To illustrate this, suppose N = 3 and the set of points where $d_{S_1} = d_{S_2}$ is not empty, and the set where $d_{S_2} = d_{S_3}$ is not empty (therefore, $d_{S_1} \neq d_{S_3}$ throughout R^3). Assume the plane $C(x_1, x_2, x_3) = C_0$ is divided as shown in Figure 2a, where the sets A_1, A_2, and A_3 do not include the intersections of the planes $d_{S_1} = d_{S_2}$ and $d_{S_1} = d_{S_3}$ with the plane $C(\bar{x}) = C_0$. If $d_{S_1} > d_{S_3}$ in A_1, and $d_{S_2} > d_{S_1}$ in A_3, then it follows that $d_{S_1} < d_{S_3}$ in $A_2 \cup A_3$ and $d_{S_2} < d_{S_1}$

in $A_2 \cup A_1$. Since $d_{S_2} \neq d_{S_3}$ in $C(\bar{x}) = C_o$, and $d_{S_3} > d_{S_1} > d_{S_2}$ in A_2, it follows that $d_{S_3} > d_{S_2}$ in $C(\bar{x}) = C_o$. Consequently, the system $(0, 0, x_{30}) > (x_1, x_2, x_3)$ for every $(x_1, x_2, x_3) \in A_2 \cup A_3$ and $(x_{10}, 0, 0) > (x_1, x_2, x_3)$ for every $(x_1, x_2, x_3) \in A_1$. Since this is true for every C_o, the maximum of D must lie on the hatched lines in Figure 2b.

6.3.1 One realization of the types of V and C functions discussed above can be found in Example 6.1.1 if the costs of the weapons used by the enemy are taken into account.

6.4 Subsystems in Series

6.4.1 Consider a system composed of x paralleled subsystems of kind A in series with y paralleled subsystems of kind B, as schematized in Figure 3.

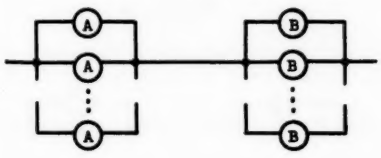

Figure 3

One realization of such a system would arise in a circuit where B represents a timer which, at a prescribed instant, allows a switch A to be actuated. A second realization relates to an air defense problem, where the assumption is made that it is possible to stop an enemy attack only if at least one radar warning device and at least one antiaircraft weapon function successfully.

Returning to the general system, suppose A has a failure probability p and a cost c and that B has a failure probability q and a cost d. If it is assumed that the value V expected to be received is proportional to the probability that the system functions and that the cost c is the sum of the component (subsystem) costs, then

$$V = V_o (1 - p^x)(1 - q^y),$$

$$C = xc + yd.$$

According to Theorem 3.1, the pair $(V_x/C_x, V_y/C_y) = [V_o(1 - q^y)p^x |\ln p|/c, V_o(1 - p^x)q^y |\ln q|/d]$ is a preference function. By Theorem 3.5, the maximum of any $D \in C^*$ lies on the curve $V_x/C_x - V_y/C_y = 0$ or on the boundary ($x = 0$ or $y = 0$ in this case), where

$$V_x/C_x - V_y/C_y = (V_o p^x q^y |\ln p|/c)[q^{-y} - \lambda(p^{-x} - 1) - 1],$$

$$\lambda = c|\ln q|/d|\ln p|.$$

If (x_o, y_o) is any point on the curve $q^{-y} - \lambda(p^{-x} - 1) - 1 = 0$, then for $(x = x_o, y > y_o)$, $V_x/C_x - V_y/C_y > 0$ and for $(x > x_o, y = y_o)$, $V_x/C_x - V_y/C_y < 0$, since this curve passes through (0, 0) and has a positive slope. Hence D increases as the curve $V_x/C_x - V_y/C_y = 0$ is approached from either side along a constant-cost or constant-value curve. This means the maximum of D lies on the curve $V_x/C_x - V_y/C_y = 0$ and not on the boundary of R^2. The character of this curve is shown in Figure 4 for the case $C_1 > C_2$ and $|\ln p| < |\ln q|$.

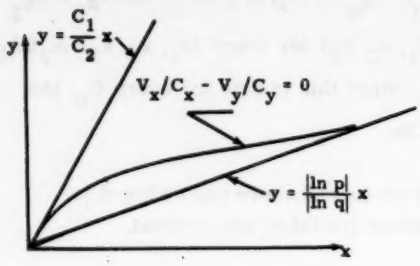

Figure 4

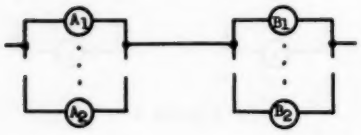

Figure 5

6.4.2 It is instructive to go a step further with this type of example and allow the designs of A and B to vary. Suppose x_1 subsystems of type A_1 and x_2 subsystems of type A_2, all in parallel, and in series with a parallel arrangement of x_3 subsystems of type B_1 and x_4 subsystems of type B_2 (see Figure 5), are introduced into the system. In the order thus established, let the failure probabilities be p_1, p_2, q_1, q_2 and the costs c_1, c_2, d_1, d_2. Under the same assumptions as made in the preceding example,

$$V = V_0 (1 - p_1^{x_1} p_2^{x_2})(1 - q_1^{x_3} q_2^{x_4}),$$

$$C = x_1 c_1 + x_2 c_2 + x_3 d_1 + x_4 d_2 .$$

Without loss in generality, it is assumed that $|\ln p_1|/c_1 > |\ln p_2|/c_2$ and $|\ln q_2|/d_2 > |\ln q_1|/d_1$. It follows that V and C belong to the class P^4 on the space R^4 of vectors $\bar{x} = (x_1, x_2, x_3, x_4)$, since the index integers 1, 2, 3, 4 can be partitioned into two sets: $(1, 2) = I_1$ and $(3, 4) = I_2$, with $F_1 = F_2 = V_0 p_1^{x_1} p_2^{x_2}(1 - q_1^{x_3} q_2^{x_4})$, $F_3 = F_4 = V_0 q_1^{x_3} q_2^{x_4}(1 - p_1^{x_1} p_2^{x_2})$, $f_1 = |\ln p_1|$, $f_2 = |\ln p_2|$, $f_3 = |\ln q_1|$, $f_4 = |\ln q_2|$, $g_1 = c_1$, $g_2 = c_2$, $g_3 = d_1$, $g_4 = d_2$, $f_1/g_1 > f_2/g_2$, and $f_4/g_4 > f_3/g_3$. Consequently, Theorem 4.6 applies, with the result that $(|\ln p_1|/c_1, |\ln p_2|/c_2)$ is an invariant preference function over the subspace of vectors $(x_1, x_2, 0, 0)$, and $(|\ln q_1|/d_1, |\ln q_2|/d_2)$ is an invariant preference function over the subspace of vectors $(0, 0, x_3, x_4)$. This means there always exist x_1 and x_4 such that $(x_1, 0, 0, x_4) > (x_1', x_2', x_3', x_4')$ for any system (x_1', x_2', x_3', x_4'). Thus, the system has been reduced to all A_1's and all B_2's, so that, according to Section 4.7 $\left(V_{x_1}/C_{x_1}, V_{x_4}/C_{x_4}\right)$ is a preference function, and one can proceed as before to obtain a relationship between x_1 and x_4, as in Figure 4.

This example brings out an interesting feature of systems corresponding to $V, C \in P^N$, namely, optimization can be carried out in each subspace (i.e., corresponding to I_n) independent of the other subspaces. In the above example, this means the designs of A and B are independent of each other and can be based on $\max(|\ln p|/c)$ and $\max(|\ln q|/d)$, respectively.

6.5 The Variables x and y Integral

In general, the assumption that x and y can assume a continuous range of values yields a solution which is not an integer pair. For many applications (in particular, those included in this paper) such a solution is physically inadmissible. Furthermore, the solution may not correspond to the nearest integer pair.

To illustrate the restrictions imposed on the various parameters by requiring x, y to be an integer pair, consider (in connection with Example 6.1.1) the particular function $D(V, C) = C$ subject to the condition $V \geq V_0$; i.e., minimize $C = (C_1x + C_2y)$ subject to $V = v_0(1 - p^x q^y) \geq V_0$, or, therefore, subject to $x|\ln p| + y|\ln q| \geq A$, where $A = |\ln(1 - V_0/v_0)|$. Geometrically (Figure 6), this means the line $C_1x + C_2y = C$ is to be moved downward and to the left, subject to the requirement that at least one point on this line shall lie above and to the right of the line $x|\ln p| + y|\ln q| = A$.

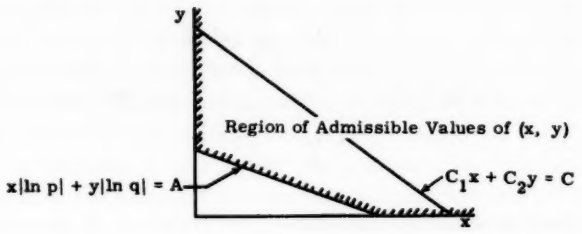

Figure 6

If x and y were continuous this would be a simple linear programming problem, the solution to which is $(x_0, y_0) = (0, A/|\ln q|)$ or $(x_0, y_0) = (A/|\ln p|, 0)$, according as $C_1/C_2 \gtrless |\ln p|/|\ln q|$. This result is the same as obtained previously in Example 6.1.1, as would be expected.

However, if the solution (x, y) is required to be an integer pair, the above analysis is valid only if $y_0 = |\ln(1 - V_0/v_0)|/|\ln q|$ is an integer, or $x_0 = |\ln(1 - V_0/v_0)|/|\ln p|$ is an integer, according as $C_1/C_2 \gtrless |\ln p|/|\ln q|$. In this connection it may be possible to adjust V_0 somewhat to make y_0 or x_0 come out integers, but the implications of such an adjustment are not easily assessed. If x_0 and y_0 are not integers, then the solution may be on an axis or on the interior of the admissible region, depending on the relative magnitude of all the parameters p, q, C_1, C_2, V_0, and v_0. The integral case will not be considered further.

6.6 Maxima and Minima of Functions

Preference functions can sometimes be used advantageously to discuss maxima and minima of functions of two or more real variables $x, y, \ldots$. As a simple example of the procedure, consider the function

$$D(x, y) = \frac{(ax + by + c)^2}{x^2 + y^2 + 1}\,; \; x > 0,\; y > 0\,.$$

So that the correspondence with the preference process shall be clear, the identifications $V = (ax + by + c)^2$ and $C = x^2 + y^2 + 1$ are made. Clearly D_C is negative, and D_V, V_x, V_y, C_x, C_y are all positive; and all functions involved and their first partial derivatives are continuous, so that it is permissible to employ the previous results.

Since

$$B(x, y) = V_x/C_x - V_y/C_y = [(ax + by + c)/xy]\,(ay - bx)\,,$$

the maximum lies on $ay = bx$, or $x = 0$ or $y = 0$. For $y > (b/a)x$, $B > 0$, so that D is increasing as the curve $ay - bx = 0$ is approached from above along, say, a constant C curve. For $y < (b/a)x$, the function $B < 0$, and D increases as $ay - bx = 0$ is approached from below, along the same curve. It follows that if D has an extreme value in R^2, it is not a minimum and it lies on the curve $ay - bx = 0$. In this way, the problem is reduced to finding the maximum of

$$D\left(x, \frac{b}{a}x\right) = \frac{\left[x\left(a + \frac{b^2}{a}\right) + c\right]^2}{x^2\left(1 + \frac{b^2}{a^2}\right) + 1}\,, \quad x > 0\,,$$

which can easily be found by standard procedure. The result is: D has a maximum of $a^2 + b^2 + c^2$ located at the point $(x, y) = (a/c, b/c)$.

There are two attractive features of this method: it may save needless computation by showing the most direct way to arrive at the solution, and it makes the sufficiency easier to establish.

7. CONCLUDING REMARKS

7.1 After the notions of preference function and invariant preference function were introduced in Sections 2.3 and 2.4, it was established that $\left(V_{x_1}/C_{x_1}, \ldots, V_{x_N}/C_{x_N}\right)$ is a preference function. Some of the interesting features of this result are: (1) this function is independent of the specific dependence of the criterion D upon V and C; (2) it is unnecessary to know the specific functions V and C, it being sufficient in every case merely to know the relations among the components of $\bar{x}$ entering in V and C; and (3) it selects from the sometimes vast number of parameters entering in V and C only those essential for the selection of a preferred system. This latter feature is particularly important since many of the parameters of D may be entirely unknown.

If V and C belong to the class of functions P^N (see Section 4.2), then the preference function $\left(V_{x_1}/C_{x_1}, \ldots, V_{x_N}/C_{x_N}\right)$ takes a much simpler form, and very specific results can be obtained (see Section 4.6, et seq.).

Two concluding remarks (Sections 7.2 and 7.3) concerning the interpretation and usefulness of the above-mentioned results are followed by some comments on open problems (Section 7.4).

7.2 Problems in system analysis frequently come encumbered with side conditions of remarkably diverse character. When this happens it is necessary to inquire into their origin and nature for the purpose of determining to what extent they can be relaxed or entirely ignored. Such an inquiry is important because the analyst must be certain that he is actually finding a best system, in some sense, rather than merely verifying a group of opinions by admitting too restrictive side conditions.

Nevertheless, it will sometimes occur, even for systems of the class discussed, that there are one or more essential side conditions. In this event, the preference function $\left(V_{x_1}/C_{x_1}, \ldots, V_{x_N}/C_{x_N}\right)$ may frequently be used to discuss the behavior of the D surface in the region R' of interest. In fact (see Theorem 3.5), it was shown for N = 2 that the maximum of D must lie either on the 'boundary' of R' or on one of the curves $V_x/C_x = V_y/C_y$.

7.3 In the case of an invariant preference function over R^N, it is always possible to decide if a system is not inferior to all others which are to be compared. But no information is obtained concerning the number of subsystems which should be employed (i.e., the number that leads to a maximum of D), because this number depends specifically on D and thus will be, in general, different for every D. If it could be shown that some D, say D_0, has greater significance than any of the others, the number of subsystems could be determined by using D_0. If such a D_0 exists, then it could be maximized from the outset, obviating the necessity for finding a preference function. However, as long as there is uncertainty about the relative merits of the various admissible D functions, it may be that the only sense in which one system can be preferred to another is in the sense of 'superior' or 'not inferior' for all criteria D, as defined herein.

7.4 The only fundamentally unrealistic assumption made was that of continuity of the components of $\bar{x}$. It would be highly desirable to obtain analogous results when the restriction to integral components of $\bar{x}$ is strictly adhered to. However, even the most simple example (that of linear programming) introduces great difficulties.

Finally, an interesting extension still within the assumption of continuity in $\bar{x}$ would be to consider the vectors $\bar{x}$ as strategies in a two-person game where the payoff is given by any member of a certain class of D functions. For example, suppose the players share a common value $V(\bar{x}, \bar{y})$ but have separate costs $C_1(\bar{x})$ and $C_2(\bar{y})$. In this case, D can be any well-behaved function of V, C_1, and C_2, which is increasing in V and C_2 and decreasing in C_1, where D is considered the payoff to the player corresponding to $\bar{x}$. If a saddle point of D exists, then theorems analogous to those proven in the foregoing can be established.

The authors wish to thank Sandia Corporation for permission to publish this paper, and Dr. R. W. Shephard for allowing them to read his paper prior to publication.

* * *

ON THE THEORY OF REPLACEMENT OF MACHINERY WITH A RANDOM FAILURE TIME[1]

George H. Weiss
U. S. Naval Ordnance Laboratory,
White Oak, Silver Spring, Maryland

This paper studies the reliability function, and moments thereof, of a machine which is periodically replaced in order to increase the mean time to failure. Two cases are studied: strictly periodic replacement and randomly periodic replacement.

I. INTRODUCTION

Some time ago N. R. Campbell [1] presented an analysis of several replacement policies for continually operating elements subject to random failures. His treatment led to several difficult integral equations. As a result, it is difficult to draw any qualitative conclusions from his work.

Specifically, Campbell considered the question of how many replacements are necessary to keep a series of street lights in continual operation. He assumed that the failure of each light could be detected instantaneously and that instantaneous replacements could be made. Under these assumptions Campbell treated two replacement policies, the first replacing the lights as they failed and the second replacing all of the lights periodically.

This paper will consider the problem from the point of view of time to first system failure with a given replacement policy. We will consider two possible replacement policies: a strictly periodic policy and a randomly periodic policy.

II. NONRANDOM PERIODIC REPLACEMENTS

We treat first the case of strictly periodic replacement. Let $R(t)$ be the reliability function, i.e., the probability that a device survives for a time t or greater. If $R(t)$ is differentiable, let $\rho(t) = -dR/dt$. Let us assume that a given mechanism is replaced every t_0 units of time and inquire for the probability density for a failure at time t.

Let $R_k(t)$ be the reliability function if k replacements have been made and no failures have occurred. Then, if $kt_0 \leq t < (k+1)t_0$, we have

---

[1] This work was performed while the author was employed at the Ballistic Research Laboratories, Aberdeen Proving Ground, Aberdeen, Maryland.

*Manuscript received March 14, 1956.

$$R_k(t) = [R(t_0)]^k \, R(t - kt_0) . \tag{1}$$

The n'th moment of this distribution, written $\langle t^n \rangle$, is given by

$$\langle t^n \rangle = - \sum_{k=0}^{\infty} \int_{kt_0}^{(k+1)t_0} t^n \, dR_k(t)$$

$$= - \sum_{k=0}^{\infty} [R(t_0)]^k \int_{kt_0}^{(k+1)t_0} t^n \, dR(t - kt_0)$$

$$= - \sum_{k=0}^{\infty} [R(t_0)]^k \int_0^{t_0} (x + kt_0)^n \, dR(x) \tag{2}$$

$$= - \sum_{k=0}^{\infty} [R(t_0)]^k \sum_{j=0}^{n} \binom{n}{j} (kt_0)^{n-j} \int_0^{t_0} x^j \, dR(x) .$$

The summation over k can be effected in closed form by manipulation of the geometric series. For n an integer we have

$$\sum_{k=0}^{\infty} k^n x^k = \left(x \frac{d}{dx} \right)^n \frac{1}{1 - x} = \frac{1}{(1 - x)} \sum_{j=1}^{n} G_n^j \; j ! \left(\frac{x}{1 - x} \right)^j , \tag{3}$$

where the G_n^j are Stirling numbers of the second kind [2]. Denote by $\langle t^j \rangle_0$ the integral

$$\langle t^j \rangle_0 = - \int_0^{t_0} x^j \, dR(x) . \tag{4}$$

Then the n'th moment $\langle t^n \rangle$ can be written

$$\langle t^n \rangle = \sum_{j=0}^{n} \binom{n}{j} t_0^{n-j} \langle t^j \rangle_0 \sum_{k=1}^{n-j} \frac{G_{n-j}^k \, k ! \, [R(t_0)]^k}{[1 - R(t_0)]^{k+1}} . \tag{5}$$

The first moment or expected time to failure is given by

$$\langle t \rangle = \frac{t_0 \, R(t_0) + \langle t \rangle_0}{1 - R(t_0)}$$

$$= \frac{\int_0^{t_0} R(t) \, dt}{1 - R(t_0)} , \tag{6}$$

and the second moment is given by

$$(7)\qquad \langle t^2\rangle = \frac{2\int_0^{t_0} t\,R(t)\,dt}{[1-R(t_0)]} + \frac{2t_0\,R(t_0)\int_0^{t_0} R(t)\,dt}{[1-R(t_0)]^2}.$$

A point of some interest is the limiting form of these moments when $t_0 \sim 0$. If we suppose that $R(t)$ has at least one derivative at $t = 0$ then the integrals in Eqs. (6) and (7) can be approximated for $t_0 \sim 0$

$$\int_0^{t_0} t\,R(t)\,dt \sim \frac{t_0^2}{2},$$

$$(8)\qquad \int_0^{t_0} R(t)\,dt \sim t_0,$$

$$1 - R(t_0) \sim \rho(t_0)\,t_0,$$

so that for $t_0 \sim 0$ we have

$$\langle t\rangle \sim \frac{1}{\rho(t_0)},$$

$$(9)\qquad \sigma^2 = \langle t^2\rangle - \langle t\rangle^2 \sim \frac{1}{[\rho(t_0)]^2}.$$

The limiting form of the first moment indicates that an indefinite increase in the expected time of first failure can be achieved by shorter and shorter replacement intervals only if the first derivative of the reliability function is zero initially.

In [3] Herd gave a brief discussion of replacement policy based solely on the criterion of failure rate.[2] He considered two cases: a monotone increasing failure rate and a monotone decreasing failure rate. When an increasing failure rate describes the mechanism a periodic replacement program is advisable, while in the contrary case a periodic replacement policy would remove from the system mechanisms with desirable failure characteristics. The value of $\langle t\rangle$ can be used in a similar fashion. If $\langle t\rangle$ is a decreasing function of t_0 then a replacement policy would be beneficial. If $\langle t\rangle$ is an increasing function of t_0 then no replacement policy is advisable. An intermediate case is also possible. Consider the reliability function

$$(10)\qquad R(t) = 1 - \left[1 - \frac{1}{(1+\alpha t)^2}\right]^2.$$

[2] Herd defines failure rate as ". . . the probability of failure in a finite interval of time, given the age of the product."

The function $\langle t \rangle$ is found to be

$$\langle t \rangle = \frac{1}{3\alpha} \left\{ \frac{(1 + \alpha t_0)(3 + 9\alpha t_0 + 5\alpha^2 t_0^2)}{\alpha t_0 (2 + \alpha t_0)} \right\}. \tag{11}$$

A plot of $3\alpha \langle t \rangle$ as a function of αt_0 is given in Figure 1. It can be seen that $\langle t \rangle$ has a minimum which occurs at $\alpha t_0 \sim 1.29$. The curve in Figure 1 has an asymptotic value equal to 5. The function $\langle t \rangle$ is equal to its limiting value for $t_0 \sim 0.33\alpha^{-1}$, so that it is clearly inadvisable to institute a replacement program unless the condition $t_0 < 0.33\alpha^{-1}$ can be met.

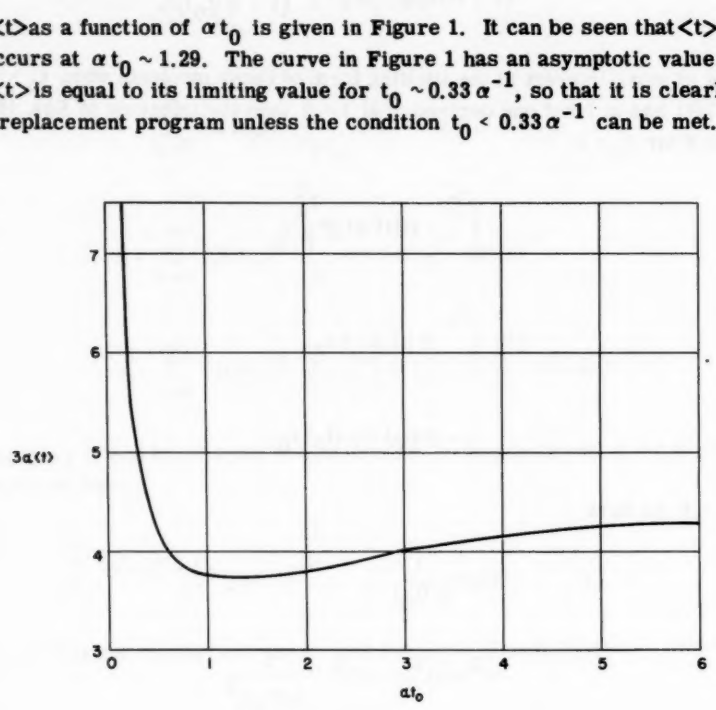

Figure 1

We can also examine this case by using a criterion derived from that of [3]. The probability that a given mechanism will fail sometime during the time interval (h, h + t), conditional upon its survival till time h, is R(t|h) where

$$R(t|h) = \frac{R(h) - R(t + h)}{R(h)}. \tag{12}$$

If we define the hazard rate, H(t), as[3] $\lim_{dt \to 0} R(dt|t)/dt$, we find

$$H(t) = -\frac{1}{R(t)} \frac{dR(t)}{dt}, \tag{13}$$

[3] This is the "force of mortality" as defined by Davis in [4].

as given by Herd. For the particular reliability function given in Eq. (10), the function $H(t)$ is given by

$$H(t) = \frac{4\alpha\,(2\alpha t + \alpha^2 t^2)}{(1+\alpha t)\,(1+4\alpha t+2\alpha^2 t^2)}. \tag{14}$$

A graph of $H(t)/4\alpha$ as a function of αt is shown in Figure 2. This function has a maximum for $t \sim 0.5\alpha^{-1}$. The hazard rate increases until t reaches this value, then decreases thereafter. Both of these criteria—that in Eq. (11), and Herd's which is essentially Eq. (14)—would suggest a replacement program if it is possible to make replacement faster than a certain rate. A more comprehensive analysis of the factors governing the choice of a reliability program would take into account the economics of the situation. A discussion of replacement program from this point of view is to be found in [5].

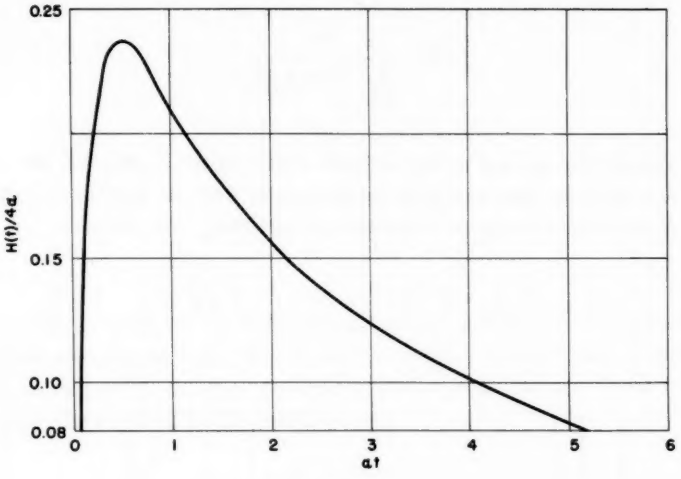

Figure 2

A particular calculation of some interest concerns the value of $\langle t \rangle$ for a network with n components in parallel, each with reliability function $R(t) = \exp(-\alpha t)$. The reliability function for the network is

$$R(t) = 1 - \left(1 - e^{-\alpha t_0}\right)^n. \tag{15}$$

A calculation carried out in the manner of Appendix A of [6] shows that

$$\alpha\langle t \rangle = \frac{B_n\left(1 - e^{-\alpha t_0}\right)}{\left(1 - e^{-\alpha t_0}\right)^n} \tag{16}$$

where the functions $B_n(x)$ are

$$B_n(x) = x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots + \frac{x^n}{n} . \tag{17}$$

For a single component we have $\alpha\langle t\rangle = 1$ so that the functions $B_n\left(1 - e^{-\alpha t_0}\right)\Big/\left(1 - e^{-\alpha t_0}\right)^n$ represent the improvement afforded by the use of a replacement program and redundant components. Several curves of $\alpha\langle t\rangle$ as a function of αt_0 are shown in Figure 3. For large n the functions $B_n(x)$ can be written

$$B_n(x) = -\log_{|x|<1}(1 - x) + o(x^n) , \tag{18}$$

so that Eq. (16) becomes, approximately,

$$\langle t\rangle \sim \frac{t_0}{\left(1 - e^{-\alpha t_0}\right)^n} . \tag{19}$$

Equation (19) indicates that when a policy of finite replacement is adopted, the mean time to first system failure goes up approximately exponentially with the number of redundant components. If there is no replacement, or equivalently $\exp(-\alpha t_0) = 0$, we have

$$B_n(1) = 1 + \frac{1}{2} + \cdots + \frac{1}{n} \sim \log n , \tag{20}$$

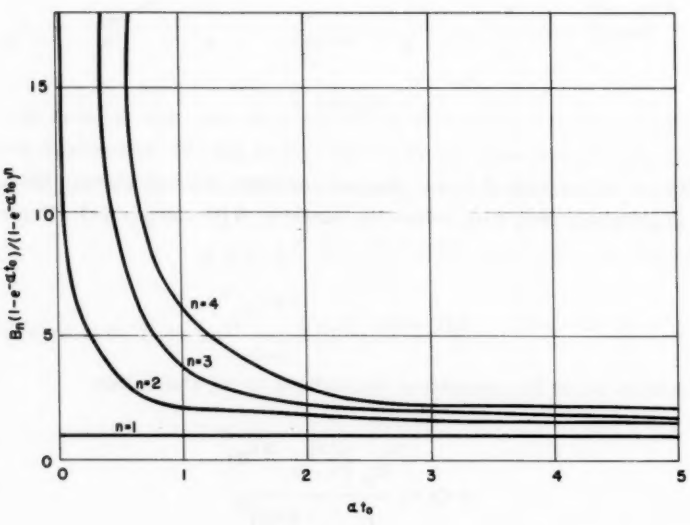

Figure 3

where the last estimate follows from the well-known asymptotic expansion of the harmonic series (cf. [7]). Hence, if no finite replacement policy is adopted, the effect of adding redundant components shows up approximately as the logarithm of the number of components.

We can also examine the distribution of the number of replacements made before a system failure occurs. The probability that a given device (or system) will function correctly for a time t_0 is $R(t_0)$. Hence, the probability that there will be n replacements before a system failure occurs is $[1 - R(t_0)]\,[R(t_0)]^n$. If the moments of this distribution are defined by

$$\langle n^k \rangle = [1 - R(t_0)] \sum_{n=0}^{\infty} n^k \, [R(t_0)]^n , \tag{21}$$

then we have

$$\langle n \rangle = \frac{R(t_0)}{1 - R(t_0)} ,$$

$$\langle n^2 \rangle = \frac{R(t_0)\,[1 + R(t_0)]}{[1 - R(t_0)]^2} , \tag{22}$$

$$\sigma^2 = \langle n^2 \rangle - \langle n \rangle^2 = \frac{R(t_0)}{[1 - R(t_0)]^2} .$$

In our previous paragraphs we have allowed the reliability function to depend only upon the time. In general, one should study the reliability function as a function of other parameters, such as temperature, humidity, or other relevant environmental factors. With this in mind we may write the reliability function in the general form,

$$R(t) = R(\beta_1, \beta_2, \ldots \beta_j ; t) \tag{23}$$

where now the β's may be random variables. Let us now define $F(x_1, x_2, \ldots x_j; t)$ as the distribution function of the β's at time t, that is,

$$F(x_1, x_2, \ldots , x_j ; t) = \Pr \{\beta_1 \le x_1, \beta_2 \le x_2, \ldots \beta_j \le x_j \text{ at time } t\} . \tag{24}$$

In order to get the values of $\langle t^n \rangle$ we may define an averaged reliability function $R^*(t)$, where

$$R^*(t) = \int_{-\infty}^{\infty} \cdots \int R(\alpha_1, \alpha_2, \ldots \alpha_n ; t)\, d_{\alpha_1 \alpha_2 \ldots \alpha_n} F(\alpha_1, \alpha_2, \ldots \alpha_n ; t) , \tag{25}$$

and then take moments with respect to $R^*(t)$ as in Eq. (2). As an illustration, let us consider a reliability function $R(t) = \exp(-\alpha t)$, where α has the probability density $k[\exp(-k\alpha)]$. The function $R^*(t)$ is

$$R^*(t) = k \int_0^\infty e^{-\alpha(k+t)}\, d\alpha = \frac{k}{k+t}, \tag{26}$$

which implies that for periodic replacement $\langle t \rangle$ is given by

$$\langle t \rangle = k\left(1 + \frac{k}{t_0}\right) \ln\left(1 + \frac{t_0}{k}\right). \tag{27}$$

We can see from Eq. (27) that even though the original reliability function is an exponential the first moment is a function of t_0.

III. RANDOM PERIODIC REPLACEMENTS

It can happen that it is impractical to replace a mechanism in a strictly periodic fashion. This might, for example, be the case when a given mechanism has a variable work cycle and is of such a nature that a replacement in midcycle is impossible or impractical. In this eventuality the replacement policy would have to be a random one, taking advantage of any free time available for making the replacements.

To describe the replacement policy we will assume that there exists a function $F(t)$ which is the probability that no replacement is made in the time interval $(0, t)$. If it is certain that some replacement will eventually be made, $F(t)$ satisfies

$$F(0) = 1, \quad F(\infty) = 0. \tag{28}$$

A further condition on $F(t)$ is that it be a monotone nonincreasing function. It is also possible to describe a periodic replacement policy in terms of $F(t)$ by taking

$$\begin{aligned} F(t) &= 1, \quad t < t_0, \\ &= 0, \quad t \geq t_0. \end{aligned} \tag{29}$$

We shall solve for a reliability function $R(t)$ analogous to that given in Eq. (11) for the case of strictly periodic replacements. Let us assume that replacements can be made instantaneously. In contrast to the situation treated in the last section, the number of replacements made in time t is a random variable. We shall show that $R(t)$ is the solution of an integral equation of the kind encountered in renewal theory.

The event "a system survives for a time t or greater" is the conjunction of events "the system survives for a time t or greater with j replacements having been made," where $j = 1, 2, 3, \ldots$. Hence $R(t)$ can be written,

$$\mathfrak{R}(t) = R(t)\,F(t) - \int_0^t dF(\tau)\,R(\tau)\,R(t-\tau)\,F(t-\tau)$$

$$+ \int_0^t dF(\tau_1) \int_{\tau_1}^t dF(\tau_2-\tau_1)\,R(\tau_1)\,R(\tau_2-\tau_1)\,R(t-\tau_2)\,F(t-\tau_2)$$

(30)

$$- \int_0^t dF(\tau_1) \int_{\tau_1}^t dF(\tau_2-\tau_1) \int_{\tau_2}^t dF(\tau_3-\tau_2)\,R(\tau_1)\,R(\tau_2-\tau_1)\,R(\tau_3-\tau_2)\,R(t-\tau_3)\,F(t-\tau_3)$$

$$+ \cdots .$$

We can simplify Eq. (30) by introducing two auxiliary functions defined by

$$\Omega(t) = R(t)\,F(t)\,,$$

(31)

$$\psi(t) = -\int_0^t R(\tau)\,dF(\tau)\,.$$

The function $\Omega(t)$ is the probability that neither a failure nor a replacement occur in time t. The function $\psi(t)$ is the probability that at least one replacement be made in time t. Equation (30) can be rewritten, by using Eq. (31),

$$\mathfrak{R}(t) = \Omega(t) + \int_0^t \Omega(t-\tau)\,d\psi(\tau)$$

$$+ \int_0^t d\psi(\tau_1) \int_0^{t-\tau_1} d\psi(\tau_2-\tau_1)\,\Omega(t-\tau_2)$$

(32)

$$+ \int_0^t d\psi(\tau_1) \int_0^{t-\tau_1} d\psi(\tau_2-\tau_1) \int_0^{t-\tau_2} d\psi(\tau_1-\tau_2)\,\Omega(t-\tau_3)$$

$$+ \cdots .$$

This expression for $\mathfrak{R}(t)$ is seen to be the Neumann series, defining the solution of

$$\mathfrak{R}(t) = \Omega(t) + \int_0^t \mathfrak{R}(t-\tau)\,d\psi(\tau)\,. \tag{33}$$

If $R(t)$ is continuous and $F(t)$ has a derivative $-g(t)$, then $\psi(t)$ is differentiable. We shall let its derivative be denoted by $\phi(t)$. Under these conditions Eq. (33) is equivalent to

$$R(t) = \Omega(t) + \int_0^t R(t-\tau)\,\phi(\tau)\,d\tau\,. \tag{34}$$

Although it is seldom possible to find an explicit solution to either of Eqs. (33) and (34), we may still find the moments of $R(t)$ by differentiating its Laplace transform at the origin. Let us define the following transforms:

$$R^*(s) = \int_0^\infty e^{-st}\,dR(t)\,,\quad \Omega^*(s) = \int_0^\infty e^{-st}\,d\Omega(t)\,,$$
$$\psi^*(s) = \int_0^\infty e^{-st}\,d\psi(t)\,. \tag{35}$$

The Laplace transform of Eq. (33) yields for $R^*(s)$ the expression

$$R^*(s) = \frac{\Omega^*(s)}{1-\psi^*(s)}\,. \tag{36}$$

The moments of $R(t)$ can now be recovered in terms of derivatives of $R^*(s)$:

$$\langle t^n\rangle = (-1)^n \left.\frac{d^n R^*(s)}{ds^n}\right|_{s=0+}\,. \tag{37}$$

In the simpler case where there exists the function $\phi(t)$ we may define another set of transforms:

$$R_1^*(s) = \int_0^\infty R(t)\,e^{-st}\,dt\,,\quad \Omega_1^*(s) = \int_0^\infty \Omega(t)\,e^{-st}\,dt\,,$$
$$\phi^*(s) = \int_0^\infty \phi(t)\,e^{-st}\,dt\,. \tag{38}$$

In this case we have

$$R_1^*(s) = \frac{\Omega_1^*(s)}{1-\phi^*(s)}\,. \tag{39}$$

The moments may be written,[4]

$$\langle t^n \rangle = -\int_0^\infty t^n \, d\bar{R}(t) = n \int_0^\infty t^{n-1} \, \bar{R}(t) \, dt$$

(40)

$$= (-1)^n \, n \, \frac{d^{n-1} \bar{R}_1^*(s)}{ds^{n-1}} .$$

When the function $\bar{R}_1^*(s)$ is used, computation of the first moment is particularly simple. As an example of the use of Eq. (40), let us consider the reliability function defined by Eq. (15), with $F(t)$ equal to $\exp(-\beta t)$. The functions $\Omega_1^*(s)$ and $\phi^*(s)$ are given by

$$\Omega_1^*(s) = \sum_{j=1}^{n} (-1)^{j+1} \binom{n}{j} \frac{1}{s + \beta + \alpha_j} ,$$

(41)

$$\phi^*(s) = \beta \, \Omega_1^*(s) .$$

For the first moment we find

$$\langle t \rangle = \frac{\Omega_1^*(0)}{1 - \beta \, \Omega_1^*(0)}$$

(42)

$$= \frac{1}{\beta} \left\{ \prod_{j=1}^{n} \left(1 + \frac{1}{j} \frac{\beta}{\alpha} \right) - 1 \right\} .$$

Details of the integration are given in the Appendix. Several curves of $\alpha \langle t \rangle$ are plotted in Figure 4 as a function of α/β.

---

[4] In deriving the second integral from the first, use was made of the fact that if $\langle t^n \rangle$ is finite then $\lim_{t \to \infty} \bar{R}(t) \, t^n = 0$. This is easily seen from the relations

$$t^n \bar{R}(t) = -t^n \int_t^\infty d\bar{R}(x) \leq -\int_t^\infty x^n \, d\bar{R}(x) .$$

The existence of a finite $\langle t^n \rangle$ insures that the last integral tends to zero for $t \to \infty$.

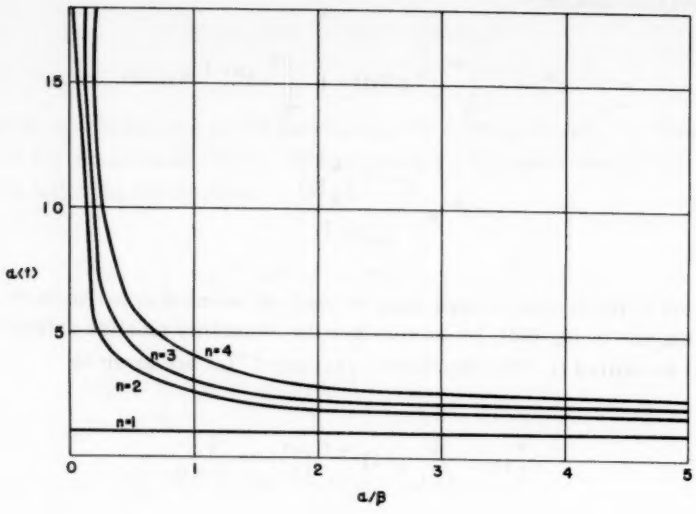

Figure 4

APPENDIX

Derivation of Eq. (42)

Equation (41) indicates that $\Omega_1^*(0)$ is given by

$$\Omega_1^*(0) = \sum_{j=1}^{n} (-1)^{j+1} \binom{n}{j} \frac{1}{\beta + \alpha j}. \tag{A1}$$

We will now evaluate this sum. Consider a function J(u) given by

$$\begin{aligned} J(u) &= \sum_{j=1}^{n} (-1)^{j+1} \binom{n}{j} u^{\beta+\alpha j-1} \\ &= u^{\beta-1} \left[1 - (1 - u^{\alpha})^n\right]. \end{aligned} \tag{A2}$$

It is easily seen that

$$\begin{aligned} \Omega_1^*(0) &= \int_0^1 J(u)\, du \\ &= \frac{1}{\beta} - \int_0^1 u^{\beta-1} (1 - u^{\alpha})^n \, du. \end{aligned} \tag{A3}$$

We change the variable of integration by setting

$$v = u^{\alpha}, \quad du = \frac{1}{\alpha} v^{\frac{1}{\alpha}-1}. \tag{A4}$$

Equation (A3) then becomes

$$\Omega_1^*(0) = \frac{1}{\beta} - \frac{1}{\alpha} \int_0^1 v^{\frac{\beta}{\alpha}-1} (1 - v)^n \, dv. \tag{A5}$$

However, the last integral is just the Beta function [8], i.e.,

$$\Omega_1^*(0) = \frac{1}{\beta} - \frac{1}{\alpha} B\left(\frac{\beta}{\alpha}, n+1\right)$$

$$= \frac{1}{\beta} - \frac{1}{\alpha} \frac{n!\,\Gamma\left(\frac{\beta}{\alpha}\right)}{\Gamma\left(n+1+\frac{\beta}{\alpha}\right)}$$

(A6)

$$= \frac{1}{\beta} - \frac{n!}{\alpha \prod_{j=0}^{n} \left(j+\frac{\beta}{\alpha}\right)},$$

$$= \frac{1}{\beta}\left\{1 - \frac{1}{\prod_{j=1}^{n}\left(1+\frac{\beta}{j\alpha}\right)}\right\},$$

where the second line is obtained from the first by using the well-known expression for the Beta function in terms of Gamma functions. Hence, we have the expression

$$\langle t \rangle = \frac{\Omega_1^*(0)}{1 - \beta\,\Omega_1^*(0)}$$

(A7)

$$= \frac{1}{\beta}\left\{\prod_{j=1}^{n}\left(1+\frac{1}{j}\frac{\beta}{\alpha}\right) - 1\right\}.$$

which was to be proved.

REFERENCES

[1] N. R. Campbell, "The Replacement of Perishable Members of a Continually Operating System," Supplement, Journal of the Royal Statistical Society 7:110 (1941).

[2] C. Jordan, Calculus of Finite Differences, Rottig and Romwalter, Sopron Hungary, 1939, first edition.

[3] G. R. Herd, "Failure Rates," Arinc Monograph 2:31 (1955).

[4] J. Davis, "An Analysis of Some Failure Data," Journal of the American Statistical Association 47:113 (1952).

[5] G. Weiss, "On Some Economic Factors Influencing a Reliability Program," NAVORD Report 4256, U. S. Naval Ordnance Laboratory, White Oak, Md.

[6] G. Weiss, "On the Time-Dependent Reliability of Networks," Proceedings of the National Electronics Conference (National Electronics Conference Inc., Chicago, 1956) II, first edition, p. 1017.

[7] K. Knopp, Theory and Application of Infinite Series, Blackie and Son, Ltd., London, England, 1951, second English edition.

[8] E. T. Copson, Theory of Functions of a Complex Variable, Oxford University Press, London, England, 1935, first edition.

* * *

AN INVENTORY CONTROL BIBLIOGRAPHY

R. Lewis, F. Neeland
and
M. Gourary
The George Washington University
Logistics Research Project

A bibliography of inventory control arranged in three sections: Technical Publications, Machine Methods and Applications, and Mathematical and Statistical Techniques.

Technical Publications

Ablahat, Newton A. An operations research approach to inventory controls (abstract). Journal of the Operations Research Society of America, Vol. 2, No. 3, August 1954, p. 346.

Ackoff, Russell L. Production and inventory control in a chemical process. Journal of the Operations Research Society of America, Vol. 3, No. 3, August 1955, pp. 319-338.

Ackoff, Russell L., and James E. Townsend. Control of production and inventory in a chemical process (abstract). Journal of the Operations Research Society of America, Vol. 2, No. 3, August 1954, pp. 347-348.

Allen, W. R. See Minden, D., and W. R. Allen.

Arrow, Kenneth J., Theodore Harris, and Jacob Marschak. Optimal inventory policy. Econometrica, Vol. 19, No. 3, July 1951, pp. 250-272.

---- ---- and ---- Some inventory models. Santa Monica, Calif., RAND Corporation, August 1950 (LOGS 106)

Barber, Joseph H. Economic control of inventory. New York, Codex Book Company, 1925.

Barrer, D. Y. A waiting line problem characterized by impatient customers and indifferent clerks (abstract). Journal of the Operations Research Society of America, Vol. 3, No. 3, August 1955, p. 360.

Bellman, R. On a functional equation arising in the problem of optimal inventory. Santa Monica, Calif., RAND Corporation, January 1954 (RAND paper No. P-480)

*Manuscript received January 5, 1956.

---- On some mathematical problems arising in the theory of optimal inventory and stock control. Santa Monica, Calif., RAND Corporation, October 1954 (RAND paper No. P-580)

Berman, Edward B., and A. J. Clark. An optimal inventory policy for a military organization. Santa Monica, Calif., RAND Corporation, March 1955 (RAND paper No. P-647)

Brown, B. B. See Youngs, J. W. T., M. A. Geisler, and B. B. Brown.

Charnes, A., and H. A. Simon. A note on the inventory-production control model with a perfect forecast. Pittsburgh, Graduate School of Industrial Administration, Carnegie Institute of Technology, March 1954 (ONR Research memorandum No. 13)

Clark, A. J. See Berman, E. B., and A. J. Clark

Clark, W. Van Allan, Jr., and William E. Ritchie. Economic lot-size and inventory control. N.A.C.A. Bulletin, Vol. 34, No. 6, February 1953, pp. 772-782.

Culbertson, Robert W. Solving production and inventory problems. Research Reviews, Office of Naval Research, December 1954, pp. 8-12.

Culbertson, Robert W., and Charles C. Holt. Production control and inventory control practices and problems as evidenced by a survey of eleven American companies. Middle Atlantic Conference Transactions, American Society for Quality Control, 1956.

Dannerstedt, Gunnar. Production scheduling for an arbitrary number of periods given the sales forecast in the form of a probability distribution. Journal of the Operations Research Society of America, Vol. 3, No. 3, August 1955, pp. 300-318.

Darling, D. A. See Kincaid, W. M., and D. A. Darling.

Darling, D. A., and W. M. Kincaid. An inventory problem (abstract). Journal of the Operations Research Society of America, Vol. 1, No. 2, February 1953, p. 80.

Davidson, J. H. The warehousing problem (abstract). Journal of the Operations Research Society of America, Vol. 1, No. 3, May 1953, p. 150.

Dickie, H. Ford. Six steps to better inventory management. General Electric Company (Reprinted from Factory Management and Maintenance, McGraw-Hill, August 1953).

Dvoretzky, A., J. Kiefer, and J. Wolfowitz. The inventory problem: I. Case of known distributions of demand; II. Case of unknown distributions of demand. Econometrica, Vol. 20, No. 2, April 1952, pp. 187-222; Vol. 20, No. 3, July 1952, pp. 450-466.

Abstract and discussions by S. S. Cairns and T. M. Whitin in Minutes of the Third Annual Logistics Conference, 1952. Washington, Logistics Research Project, The George Washington University, 1952.

---- ---- and ---- On the optimal character of the (s,S) policy in inventory theory. Econometrica, Vol. 21, No. 4, October 1953, pp. 586-596.

Eisenhart, Churchill. Some inventory problems. Washington, National Bureau of Standards, January 1948 (Techniques of statistical inference. Lecture 1).

---- ---- Logistics papers, issue No. 1. Washington, Logistics Research Project, The George Washington University, 1950.

Fabian, Tibor, and Ross Nelson. Production rates and inventories. University of California, Industrial Logistics Research Project, August 1953 (Research Report No. 22, pp. 1-12)

Feeney, George J., and Larry E. Hoyt. A basis for strategic decision on inventory control operations. Management Science, November 1955.

Frank, A. R. See Tuttle, C. R. M., and A. R. Frank.

Friedman, Joyce. Optimal inventories and preferred systems. Logistics papers, issue No. 5. Washington, Logistics Research Project, The George Washington University, 1951.

Geisler, M. A. See Youngs, J. W. T., M. A. Geisler, and B. B. Brown.

Geisler, M. A. See Youngs, J. W. T., M. A. Geisler, and A. R. Mirkovich.

George Washington University, Logistics Research Project. (See Logistics Research Project, The George Washington University)

Haack, Louise B. Selected supplementary bibliography of publications on inventory control. Logistics papers, issue No. 2. Washington, Logistics Research Project, The George Washington University, 1950.

Harris, Theodore. See Arrow, K. J., T. Harris, and J. Marschak.

Hohn, Franz E. See Modigliani, F., and F. E. Hohn.

Holley, Julian L. A stochastic inventory model. Alexandria, Va., Melpar, Inc., April 1954.

Holt, Charles C. Superposition decision rules for production and inventory control. Pittsburgh, Graduate School of Industrial Administration, Carnegie Institute of Technology, October 1953. (ONR Research memorandum No. 3)

Holt, Charles C. See Culbertson R. W., and C. C. Holt.

Holt, Charles C. See Simon, H. A., and C. C. Holt.

Holt, Charles, C., and Herbert A. Simon. Optimal decision rules for production and inventory control. Proceedings of the Conference on Operations Research in Production and Inventory Control, January 20-22, 1954. Cleveland, Case Institute of Technology, 1954.

Hoyt, Larry E. See Feeney, G. J., and L. E. Hoyt.

Hughes, Harry M. See Kahn, L. B., and H. M. Hughes.

Hurni, Mel. The use of operations research in inventory control. Proceedings of the Conference on Operations Research, January 1954. New York, Society for Advancement of Management.

Kahn, Louis B., and Harry M. Hughes. Another look at the inventory problem. Washington, Logistics Research Project, The George Washington University, April 1955. (Mimeographed)

Kiefer, J. See Dvoretzky, A., J. Kiefer, and J. Wolfowitz.

Kincaid, W. M. See Darling, D. A., and W. M. Kincaid.

Kincaid, W. M., and D. A. Darling. A type of inventory problem. Ann Arbor, Engineering Research Institute, University of Michigan, May 1952. (Report R 33)

Kruskal, Joseph B., Jr., and J. Jay Wolf. Consumption as a random variable. Logistics papers, issue No. 2. Washington, Logistics Research Project, The George Washington University, 1950.

Laderman, J., S. B. Littauer, and Lionel Weiss. The inventory problem. Journal of the American Statistical Association, Vol. 48, No. 264, December 1953, pp. 717-732.

Littauer, S. B. See Laderman, J., S. B. Littauer, and L. Weiss.

Little, John D. C. The use of storage water in a hydroelectric system. Journal of the Operations Research Society of America, Vol. 3, No. 2, May 1955, pp. 187-197.

Logistics Research Project, The George Washington University. Bibliography of publications pertaining to inventory control. Logistics papers, issue No. 1, 1950. (Minutes of a Working Conference on Mathematical Problems in Logistics, Appendix III)

---- The inventory control problem. Logistics papers, issue No. 1, 1950. (Minutes of a Working Conference on Mathematical Problems in Logistics, Section 4)

---- Lead time and optimal allowances: an extreme example. Logistics papers, issue No. 1, 1950. (Minutes of a Working Conference on Mathematical Problems in Logistics, Section 15)

Marschak, Jacob. Optimal aggregation of inventories under certainty. Cowles Commission Discussion paper, Economics No. 2015, March 1951. (Hectographed)

---- Optimal inventories and penalty functions. Santa Monica, Calif., RAND Corporation, August 1950. (LOGS 91)

---- Some aspects of inventory control (abstract). Econometrica, Vol. 20, No. 3, July 1952, p. 499.

Marschak, Jacob. See Arrow, K. J., T. Harris, and J. Marschak.

Mellon, B. Inventory and production scheduling by reference to cost surrogates—an example (abstract). Management Science, Vol. 1, No. 2, January 1955, p. 192.

Minden, Dorothea, and W. R. Allen. An application of waiting line theory to a simple spare parts inventory problem. Dynamic system studies: Operation and maintenance procedures for analog computers, Appendix 1. Wright Air Development Center, Wright-Patterson Air Force Base, August 1954. (WADC Technical report 54-250, Part 6)

Mirkovich, A. R. See Youngs, J. W. T., M. A. Geisler, and A. R. Mirkovich.

Modigliani, Franco. Suggestions for attacking some production and inventory problems of a factory-warehouse system (The Viscous Mfg. Co.). Pittsburgh, Graduate School of Industrial Administration, Carnegie Institute of Technology. (ONR Research memorandum No. 1).

Modigliani, Franco, and Franz E. Hohn. Production planning over time and the nature of the expectation and planning horizon. Econometrica, Vol. 23, No. 1, January 1955, pp. 46-66.

Modigliani, Franco, and John F. Muth. Optimum lot size under uncertainty and joint costs. Part 1: Certainty; Part II: Uncertainty. Pittsburgh, Graduate School of Industrial Administration, Carnegie Institute of Technology, December 1953, July 1954. (ONR Research Memorandum No. 6, No. 21)

Morin, François. Note on an inventory problem discussed by Modigliani and Hohn. Chicago, Cowles Commission for Research in Economics, University of Chicago, May 1952. (Cowles Commission Discussion paper: Economics No. 2042)

Muth, John F. See Modigliani, F., and J. F. Muth.

Naddor, Eliezer. Lead time considerations in a simplified mathematical model of inventory (abstract). Journal of the Operations Research Society of America, Vol. 3, No. 3, August 1955, p. 358.

---- Mathematical models of inventory control and their application (abstract). Management Science, Vol. 1, No. 2, January 1955, p. 191.

---- Simplified mathematical models of inventory (abstract). Journal of the Operations Research Society of America, Vol. 3, No. 1, February 1955, p. 116.

Nelson, Ross. See Fabian, T., and R. Nelson.

Ritchie, William E. See Clark, W. V., Jr., and W. E. Ritchie.

Rosenblatt, Murray. An inventory problem. Econometrica, Vol. 22, No. 2, April 1954, pp. 244-247.

Schneider, Erich. Absatz, Produktion und Lagerhaltung bei einfacher Produktion. Archiv fur Mathematische Wirtschafts und Sozialforschung, band IV, heft 1, Leipzig, 1938.

Simon, Herbert A. Dynamic programming under uncertainty with a quadratic criterion function. Econometrica, Vol. 24, No. 1, January 1956, pp. 74-81

---- On the application of servomechanism theory in the study of production control. Econometrica, Vol. 20, No. 2, April 1952, pp. 247-268.

---- Production and inventory control in the Viscous Manufacturing Company: the production rate problem. Pittsburgh, Graduate School of Industrial Administration, Carnegie Institute of Technology, November 1953. (ONR Research memorandum No. 4)

Simon, Herbert A. See Charnes, A., and H. A. Simon.

Simon, Herbert A., and Charles C. Holt. The control of inventories and production rates, a survey. Journal of the Operations Research Society of America, Vol. 2, No. 3, August 1954, pp. 289-301.

Solomon, Morris J. Optimum operation of a complex activity under conditions of uncertainty. Journal of the Operations Research Society of America, Vol. 2, No. 4, November 1954, pp. 419-432.

Steers, Newton I., Jr. A past system of inventory control. N.A.C.A. Bulletin, Vol. 33, No. 6, February 1952, pp. 753-760.

Stillson, Paul. Production and inventory control in a chemical process, I. Proceedings of the Conference on Operations Research in Production and Inventory Control, January 20-22, 1954. Cleveland, Case Institute of Technology, 1954.

Tompkins, C. B. Allocation between assignments to planned consumption and to safety allowance, a toy problem. Logistics papers, issue No. 3, Washington, Logistics Research Project, The George Washington University, 1950.

---- Determination of a safety allowance. Logistics papers, issue No. 5. Arlington, Va., Engineering Research Associates, Inc., 1949.

Townsend, James E. See Ackoff, R. L., and J. E. Townsend.

Tuttle, C. R. M., and A. R. Frank. Inventory control methods applied to electron tubes. Logistics papers, issue No. 8, I. Washington, Logistics Research Project, The George Washington University, 1952.

Tuttle, C. R. M., and J. Jay Wolf. Models of inventory control systems. Logistics papers, issue No. 8, I. Washington, Logistics Research Project, The George Washington University, 1952.

Vassian, Herbert J. Application of discrete variable servo theory to inventory control. Journal of the Operations Research Society of America, Vol. 3, No. 3, August 1955, pp. 272-282.

Vazsonyi, Andrew. The use of mathematics in production and inventory control, I: Theory of parts listing. II: Theory of scheduling. Management Science, Vol. 1, No. 1, October 1954, pp. 70-85; Vol. 1, No. 3-4, April-July 1955, pp. 207-223.

Weiss, Lionel. See Laderman J., S. B. Littauer, and L. Weiss.

---- Bibliography of publications pertaining to inventory control—supplement, Logistics papers, issue No. 1. Washington, Logistics Research Project, The George Washington University, 1950.

---- Erich Schneider's inventory control analysis. Journal of the Operations Research Society of America, Vol. 2, No. 3, August 1954, pp. 329-334.

Whitin, Thomson M. The interaction of economic purchase quantities and reorder points. Logistics papers, issue No. 4, I. Washington, Logistics Research Project, The George Washington University, 1951.

---- Inventory control in theory and practice. Quarterly Journal of Economics, November 1952, pp. 502-521.

---- Inventory control research: a survey. Management Science, Vol. 1, No. 1, October 1954, pp. 32-40.

---- Optimal inventory levels and delivery time (abstract). Journal of the Operations Research Society of America, Vol. 3, No. 3, August 1955, p. 357.

---- The theory of inventory management. Princeton, N. J., Princeton University Press, 1953. Reviewed by Alvin Karchere in Journal of the Operations Research Society of America, Vol. 1, No. 5, November 1953, pp. 314-316; by Robert M. Solow in Econometrica, Vol. 22, No. 2, April 1954, pp. 252-253; and by Thomas V. V. Atwater, Jr. in Naval Research Logistics Quarterly, Vol. 1, No. 4, December 1954, pp. 295-300.

Wolf, J. Jay. A comparison of model inventory control systems. Logistics papers, issue No. 7. Washington, Logistics Research Project, The George Washington University, 1951.

Wolf, J. Jay. See Kruskal, J. B., Jr., and J. J. Wolf.

Wolf, J. Jay, and Joseph B. Kruskal, Jr. Protection from depletion of supply in the Bureau of Medicine and Surgery. Logistics papers, issue No. 2. Washington, Logistics Research Project, The George Washington University, 1950.

Wolf, J. Jay. See Tuttle, C. R. M., and J. J. Wolf.

Wolfowitz, J. See Dvoretzky, A., J. Kiefer, and J. Wolfowitz.

Youngs, J. W. T., M. A. Geisler, and B. B. Brown. The prediction of demand for aircraft spare parts using the method of conditional probabilities. Santa Monica, Calif., RAND Corporation, January 1955. (Research Memorandum RM-1413)

Youngs, J. W. T., M. A. Geisler, and A. R. Mirkovich. Confidence intervals for Poisson parameters in logistics research. Santa Monica, Calif., RAND Corporation, September 1954. (Research Memorandum RM-1357)

Machine Methods and Applications

Battelle Memorial Institute. The application of digital data-handling equipment (MIFD) to the AF inventory control problem. September 1953.

Beals, R. P. Application of an electronic data processing machine to inventory control (abstract). Computers and Automation, Vol. 4, No. 11, November 1955, p. 32.

Bellman, R. Computational problems in the theory of dynamic programming. Santa Monica, Calif., RAND Corporation, August 1953. (RAND paper No. P-423)

Day, L. W. Production control by electronics. Systems, Vol. 19, No. 4, July-August 1955, p. 22-23.

DeCarlo, C. R. Application of electronic computing machines to operations research problems. Journal of the Operations Research Society of America, Vol. 2, No. 3, August 1954, p. 348.

---- The use of automatic and semi-automatic processing equipment in production and inventory control. Proceedings of the Conference on Operations Research in Production and Inventory Control, January 20-22, 1954. Cleveland, Case Institute of Technology, 1954.

Foster, Joe M., and Viola D. Hovsepian. A plan for processing the Ships Parts Control Center inventory control on the UNIVAC. Washington, Navy Dept., David W. Taylor Model Basin, June 1955. (Report 916)

Gainen, Leon. See Groelinger, Herbert J.

Groelinger, Herbert J. Inventory control under the program-usage-replenishment system defined for electronic processing. Washington, National Bureau of Standards, August 1954. (NBS report 3651) (Prepared for a Bureau of Supplies and Accounts project for limited distribution)

Groelinger, Herbert J., and Leon Gainen. Planning and preparation for programming an inventory control procedure. Washington, National Bureau of Standards, August 1954. (NBS report 3785) (Prepared for a Bureau of Supplies and Accounts project for limited distribution)

Harper, W. F., and W. J. McGinnity. A completely mechanized material control system. N.A.C.A. Bulletin, Vol. 31, No. 11, July 1950, pp. 1371-1377.

Hovsepian, Viola D. See Foster, Joe M., and Viola D. Hovsepian.

Macdonald, Neil. A big inventory problem and the IBM 702. Computers and Automation, Vol. 4, No. 9, September 1955, pp. 6-12, 38.

McGinnity, W. J. See Harper, W. F., and W. J. McGinnity.

Mitchell, Herbert F. Electronic computers in inventory control. Proceedings of the Conference on Operations Research in Production and Inventory Control, January 20-22, 1954. Cleveland, Case Institute of Technology, 1954.

---- Electronic computers in inventory control. Proceedings of the Conference on Operations Research, January 28 and 29, 1954. New York, Society for Advancement of Management, 1954.

Ridgway, A. O. Solving supply control problems on the IBM card-programmed electronic calculator (abstract). Computers and Automation, Vol. 4, No. 4, April 1955, p. 22.

Russell, M. E. A UNIVAC system of material control. Computers and Automation, Vol. 4, No. 4, April 1955, p. 15.

Schlain, Edward L. Application of ORDFIAC to a logistics problem (abstract). Econometrica, Vol. 22, No. 4, October 1954, p. 520.

U. S. Industrial College of the Armed Forces. Use of mechanical and electronic devices and mathematical techniques in the determination of material requirements. December 1949.

Mathematical and Statistical Techniques

Arrow, K. J., D. Blackwell, and M. A. Girshick. Bayes and minimax solutions of sequential decision problems. Econometrica, Vol. 17, Nos. 3-4, July-October 1949, pp. 214-244.

Bellman, Richard. An introduction to the theory of dynamic programming. Santa Monica, Calif., RAND Corporation, 1953. (RAND Research Memorandum R-245)

---- On some applications of the theory of dynamic programming to logistics. Naval Research Logistics Quarterly, Vol. 1, No. 2, June 1954, pp. 141-153.

---- Some functional equations in the theory of dynamic programming. Proceedings of the National Academy of Sciences, Vol. 39, No. 10, October 1953, pp. 1077-1082.

---- Some problems in the theory of dynamic programming. Econometrica, Vol. 22, No. 1, January 1954, pp. 37-48.

Blackwell, D. See Arrow, K. J., D. Blackwell, and M. A. Girshick.

Brown, A. W. A note on the use of a Pearson type III function in renewal theory. Annals of Mathematical Statistics, Vol. 11, No. 4, December 1940, pp. 448-453.

Campbell, N. R. The replacement of perishable members of a continually operating system. Supplement to the Journal of the Royal Statistical Society, Vol. 7, 1941, pp. 110-130.

Charnes, A., and C. E. Lemke. Computational theory of linear programming. II: Minimization of non-linear separable convex functionals. Naval Research Logistics Quarterly, Vol. 1, No. 4, December 1954, pp. 301-312.

Clarke, A. Bruce. The time dependent waiting line problem. Ann Arbor, Engineering Research Institute, University of Michigan, March 1953. (Report No. M720-1, R 39)

Feller, Willy. On the integral equation of renewal theory. Annals of Mathematical Statistics, Vol. 12, No. 3, September 1941, pp. 243-267.

Foulkes, J. D. Problems in combinatory analysis suggested by production scheduling, Brown University, Division of Applied Mathematics.

Girshick, M. A. See Arrow, K. J., D. Blackwell, and M. A. Girshick.

Holt, Charles C., Franco Modigliani, and Herbert A. Simon. A linear decision rule for production and employment scheduling. Management Science, Vol. II, No. 1, October 1955.

Holt, Charles C. See Simon, H. A., C. C. Holt, and F. Modigliani.

Lemke, C. E. See Charnes, A., and C. E. Lemke.

Lotka, Alfred J. A contribution to the theory of self-renewing aggregates, with special reference to industrial replacement. Annals of Mathematical Statistics, Vol. 10, No. 1, March 1939, pp. 1-25.

---- The theory of industrial replacement. Skandinavisk Aktuarietidskrift, 1940, pp. 1-14.

Matsuda, Takehiko. See Modigliani, F., and T. Matsuda.

Modigliani, Franco. A method for inverting the matrix arising from a quadratic decision criterion. Pittsburgh, Graduate School of Industrial Administration, Carnegie Institute of Technology, May 1954. (ONR Research memorandum No. 15)

Modigliani, Franco, and Takehiko Matsuda. Derivation of a linear decision rule for production and employment. Pittsburgh, Graduate School of Industrial Administration, Carnegie Institute of Technology, August 1954. (ONR Research Memorandum No. 22). Published in part in Management Science, Vol. II, No. 2, January 1956.

Modigliani, Franco. See Holt, C. C., F. Modigliani, and H. A. Simon.

Modigliani, Franco. See Simon, H., C. C. Holt, and F. Modigliani.

Preinreich, G. A. D. The present status of renewal theory. Baltimore, Waverly Press, 1940.

---- The theory of industrial replacement. Skandinavisk Aktuarietidskrift, 1939, pp. 1-19.

Simon, Herbert A. See Holt, C. C., F. Modigliani, and H. A. Simon.

Simon, Herbert A., Charles C. Holt, and Franco Modigliani. Some techniques for the solution of dynamic programming problems in production scheduling. Pittsburgh, Graduate School of Industrial Administration, Carnegie Institute of Technology, March 1955. (ONR Research memorandum No. 29)

Smith, L. Wheaton. Experience in scheduling by means of dynamic programming, General Electric, Richland, Washington, paper presented at the Third Annual Meeting of the Operations Research Society of America, June 1955.

* * *

NEWS AND MEMORANDA

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics.

Dr. I. Heller, while on a leave of absence from the George Washington University Logistics Research Project, has accepted a position as Mathematical Consultant with the Navy Management Office.

The establishment of the Army Logistics Management Center at Fort Lee, Virginia, has been announced. The new Center represents an expansion of the Army Supply Management Course established in 1954 and will, in addition to the present course, offer four new management courses in the specific fields of Procurement, Requirements, Storage and Distribution, and Maintenance, with a basis orientation in the over-all supply system.

The Secretary of the Navy has authorized the establishment of a new supply demand control point, the U. S. Navy Training Device Supply Office, effective 1 July 1956. The new SDCP is located at the U. S. Naval Training Device Center, Port Washington, New York, and will exercise inventory management over repair parts required to support all training devices, and over minor training devices, aids, and components as specifically assigned. The TDSO, commissioned on 2 July 1956, will supersede the Program Support Department of the U. S. Naval Training Device Center as the SDCP for cognizance symbol T material. The Center will continue to exercise technical control over this material.

THE INDUSTRIAL COLLEGE OF THE ARMED FORCES

The Army Industrial College was first established in 1924. In 1946 it was reconstituted as the Industrial College of the Armed Forces, under the supervision of the Joint Chiefs of Staff. The College, located at Fort McNair in Washington, D. C., conducts three closely related courses.

One of these is a 10-month resident course for senior military officers and civilian employees of the Federal Government.

The second is a correspondence course called "Emergency Management of the National Economy." This course is open to qualified officers, both Regular and Reserve, and to civilians who have attended the National Resources Conference. The correspondence course normally takes a year to finish. Reserve officers receive 48 points toward retention, promotion, and retirement if they complete the course.

The third course is the National Resources Conference, which seeks to foster a better understanding of the many interrelated problems associated with national security, a stronger appreciation of the inseparable nature of the civilian-military team, and a clear recognition of the responsibilities and capabilities of each half of the team. It seeks to encourage application of the highest level of skill, intelligence, and imagination to the complex problems which must be solved if our heritage of freedom is to be secure.

The Industrial College is unique. It is the only military school which is principally concerned with the civilian economy and civilian-military relationships. All three Industrial College courses are based on the concept that, under our democratic form of government, overall control of the military operations must be supplied by a civilian-controlled industrial complex; that moral, social, political, industrial psychological, and economic factors have a profound influence on military decisions. Both halves of the civilian-military team must understand these interrelated factors.

The Industrial College is dedicated to the achievement of closer cooperation, more harmony, better understanding, greater sympathy, and smoother functioning of the civilian-military defense team.

The College is neither a planning nor an operating agency. It is an educational institution whose purpose is to expand the nucleus of informed and thinking individuals who may be called upon to play important roles in the planning and managing of our national defense. Members of the College are not spokesmen for the Department of Defense or the military services. Their mission is to present some of the complex national security problems, many of which are current and require solution now. They present no "canned" solutions, but rather some of the factors which must be considered in arriving at solutions.

* * *

RECENT PUBLICATIONS

THE SIGNAL CORPS: THE EMERGENCY. By Dulany Terrett, Office of the Chief of Military History, Department of the Army, Superintendent of Documents, Government Printing Office, Washington 25, D. C., 354 pp.

"The more mobile an armed force becomes, the more rugged the terrain it encounters, or the more widely the force is deployed, the greater becomes the difficulty of securing and maintaining rapid, completely linked communications," General Orlando Ward points out in his Foreword to THE EMERGENCY, the first volume in the history of the Signal Corps in the Technical Services subseries and thirty-second in the series UNITED STATES ARMY IN WORLD WAR II. He describes the Signal Corps as the agency which has, as part of its mission, the responsibility of developing, procuring, and furnishing signal equipment—no mean task in an age of swift and startling progress in electronics.

The development of radar, second only to nuclear fission as the greatest scientific advance of the war, is the most important story in THE EMERGENCY. Here, for the first time, is given a complete account of the Signal Corps' part in the development of army radar. The volume also covers the development of frequency modulation and describes how, thanks to the Signal Corps, the U. S. armored force was the first to have FM, which all but revolutionized the use of tanks in World War II. The advance of crystal control is an episode linked with the FM story.

The close relationship between the communications industry and the Signal Corps in the manufacturing of equipment and in the selection and training of officers and signalmen is described, as well as the wider but very much weaker relationship between the United States and its allies, especially the British. In connection with the latter, the mission of Sir Henry Tizard and other electronic scientists and physicists to the United States in 1940 is covered.

Part One of THE EMERGENCY narrates the history of the Signal Corps from its earliest days up to World War II. It describes the Corps in World War I, postwar curtailment and peacetime procurement planning, the scope of the Signal Corps during this period, and early research in wire, radio, and radar. Part Two continues the story up to the eve of Pearl Harbor and covers the rapid expansion of the Signal Corps during the Limited and Unlimited Emergencies. The remarkable developments in the electronics field are described, as are the various Services—Pigeon, Photographic, Tactical and Administrative, and Supply—of the Corps.

A thirty-six-page appendix of World War II Signal Corps equipment will prove of interest to the communications expert and amateur alike.

THE TRANSPORTATION CORPS: MOVEMENTS, TRAINING, AND SUPPLY. By Chester Wardlow, Office of the Chief of Military History, Department of the Army, Superintendent of Documents, Government Printing Office, Washington 25, D. C.

In this volume of the series U. S. **ARMY IN WORLD WAR II**, emphasis is given to the fact that, in a great conflict fought on foreign soil, victory depends on how many soldiers and how much materiel can be moved to the oversea commands and on the timeliness of their arrival.

It becomes apparent from this history that supplying the oversea theaters involved far more than the transportation of huge tonnages. (The Army shipped about 132,000,000 measurement tons from U. S. ports during the war.) The task was an intricate one of coordination and synchronization. The right proportion had to be maintained between troopship and cargo ship capacities. Care had to be taken to ship to the theaters the types of troops and the items of supply that they needed and to deliver them when they were wanted and at ports where they could be accommodated. To assemble and ship related supply items received from many different depots and factories required careful identification and sorting at the ports of embarkation. The great variety in weight and size of military supplies made it impossible to get well-balanced cargoes and efficient use of cargo space, and the author observes that, ". . . the tendency of the commanders of active theaters, fighting thousands of miles from their main source of supply, was to place heavy requisitions in order to safeguard themselves against unforeseen supply requirements or an interruption in the line of communications . . . some oversea commanders accumulated larger stocks than they could properly administer . . . . This situation, coupled with the willingness of the theaters to detain vessels for long periods and use them as floating warehouses, led the Chief of Transportation to assert that the accumulation of unnecessarily large stocks overseas was one of the logistical mistakes of the war."

The movement of troops and equipment within the United States is an integral part of the transportation story told in this volume, and the author describes how the Military Transportation Section of the Association of American Railroads, set up early in the emergency, functioned side by side with the Chief of Transportation's staff in working out efficient routings, conserving equipment, and maintaining schedules.

The relationship of the Transportation Corps with the Office of Defense Transportation, the wartime agency charged with the control of inland transportation, and with the War Shipping Administration, which controlled the bulk of merchant shipping, comes in for careful scrutiny. The book also describes the Corps' struggle to resist attempts by powerful elements in the Army Service Forces to take over functions that had been assigned to the Transportation Corps.

While the movement of troops and materiel was the distinctive function of the Chief of Transportation, he had other exacting responsibilities. Such tasks as training troops to operate

ports, railroads, small boats, and amphibious trucks and to handle other transportation activities in the overseas commands became increasingly heavy as the war progressed. Designing and procuring the many types of small vessels required by the Army and procuring locomotives and rolling stock for the military railroads and for lend-lease were additional major concerns of the Chief of Transportation. A full account of these aspects of the transportation job appears in the volume.

MEASURING BUSINESS CHANGES. By Richard M. Snyder, John Wiley and Sons, Inc., New York, 1955.

The purpose of this book is to serve as a handbook of significant business indicators or indexes, it consists of a general description of the indicators, plus the details included in them. The indicators discussed are: (1) national income and product, (2) population, (3) labor: employment and earnings, (4) commodity prices, (5) production and business activity, (6) construction activity and costs, (7) trade, (8) financial activity, and (9) stock prices. The indicators described are primarily those published by government agencies.

MAINSPRINGS OF THE GERMAN REVIVAL. By Henry C. Wallich, Yale University Press, New Haven, 1955.

Professor Wallich views the factors which have contributed to the economic recovery and growth of West Germany since World War II. He considers the domestic economic and political policies, as well as the effects of the economic and political events which occurred in the outside world during the post-war period. One of the most interesting aspects of the work is the author's account of the economic recovery and growth through the existence of a relatively free market policy and the government's "tight money" and balanced budget programs, while most other nations were proceeding in other directions to foster economic development.

THE ECONOMICS OF NATIONAL AND MILITARY STRATEGY. By Robert E. Kuenne, United States Naval War College, 1956.

This book is a compilation of seventeen lectures delivered by the author in 1956 in implementing the economic portion of National Security Studies at the Naval War College. The lectures are,therefore, structured, with the intent of developing a framework for the analysis of economic problems arising from the planning and implementation of foreign policy. As a result, the book has the aims of a textbook in a field in which few attempts have been made to bring the tools of modern economic analysis to bear.

In Part I-- termed the "Economic Factors"--a framework is developed for the appraisal of an economy's ability to support an ambitious foreign strategy. Ten such "variables" in the strategic equation are isolated, an attempt is made to indicate their role and relative strength in the equation, and five of them are set apart for more detailed study within the context of the present world power struggle. These latter factors are the level of national product, the rate

of growth of national product, the allocation of resources, the spatial and temporal structures of an economy, and the nature and degree of external economic linkages of national economies.

Part II constitutes an attempt by the author to judge the present levels of these five variables for the United States economy, and Part III is an attempt to answer the same questions for the Soviet economy, Having presented in Part I a framework relating strategic variables of an economic character to the development of strategy, Parts II and III try to "plug in" the "values' of these five most important variables into the strategic equation in order to obtain some idea of present and future capabilities of the two major power blocs.

The author attempts to demonstrate the manner in which these five variables can be manipulated in periods of cold, limited, and total war in order to achieve desired strategic capabilities. This analysis is contained in Part IV--"Economic and Strategic Thinking"-- and includes the discussions of tax policy, deficit speding, mobilization problems, the use of input-output and linear programming techniques in such manipulation, economic warfare in the conventional sense, and foreign aid as a strategic instrument.

Lastly, in Part V, problems of economic development and their meanings in terms of shaping strategy are discussed in specific terms of the conflicting economic programs of India and China.

The author is assistant professor economics, Princeton University, who served as consultant to the Naval War College in 1955 and 1956.

MODERN MATHEMATICS FOR THE ENGINEER; Edited by Edwin F. Beckenbach, McGraw-Hill Book Company, Inc., 1957.

This book gives a good quick look at what is new and exciting in the world of highbrow mathematics doing lowbrow chores. Opening with nonlinear mechanics and closing with high-speed computers, it touches all the high points along the way – such as calculus of variations, partial differential equations, theories of prediction, games, operations research, and computing methods. There are nineteen papers arranged in three groups: Mathematical Models, Probabilistic Problems, and Computational Considerations. Each is written by an authority (frequently the authority) on the subject covered. This book will be valuable to the engineer who is opening the literature on a new subject. The papers themselves plus the excellent reference lists will take him at once into what is new and important. With the compression required to cover this wide range in about 500 pages, it is clear that the treatment is not exhaustive, but even so, it is likely that many readers will find new material and ideas in subjects they have thought they knew well.

FASTER, FASTER. By W. J. Eckert and Rebecca Jones, McGraw-Hill Book Co., Inc., 1957.

"A simple description of a Giant Electronic Calculator and the problems it solves."

Definitely recommended reading for everyone who has an interest in computing machines. The authors do not tell you how to design and build a computer but they tell you why computers

are designed and built as they are. They do not give a course in programming and coding but they describe how a problem is programmed and coded. And most important, they devote one chapter to the type of problems appropriate to computer calculation. For too long the subject of computers has been covered by monographs either too technical or too elementary for the average reader. In Faster, Faster the authors have hit a nice balance in their presentation of material and this book should be on the bookshelf of every mathematician.

* * *

AUTHOR INDEX FOR VOLUME III

	Pages
BECKMANN, M. J., LADERMAN, Jack, "A Bound on the Use of Ineffecient Indivisible Units"	245-252
BELLMAN, Richard, "Notes on the Theory of Dynamic Programming. IV - Maximization Over Discrete Sets"	67- 70
CARNEY, Adm. Robert B. (Retired), "Some Observations and Experiences in Logistics"	1- 10
FENNELL, Joseph, OSHIRO, Seiki, "The Dynamics of Overhaul and Replenishment Systems for Large Equipments"	19- 43
FRANK, Marguerite, WOLFE, Philip, "An Algorithm for Quadratic Programming"	95-110
GALE, David, "The Basic Theorems of Real Linear Equations, Inequalities, Linear Programming, and Game Theory"	193-200
GOURARY, Mina Haskind, "An Optimum Allowance List Model"	177-191
GOURARY, M., LEWIS, R. NEELAND, F., "An Inventory Control Bibliography"	295-303
HALE, Jack K., REED, Ronald L., "A Formulation of the Decision Problem for a Class of Systems"	259-277
ISBELL, J. R., MARLOW, W. H., "Attrition Games"	71- 94
JACKSON, James R., "An Extension of Johnson's Results on Job Lot Scheduling"	201-203
KIEFER, J., WOLFOWITZ, J., "Sequential Tests of Hypotheses About the Mean Occurrence Time of a Continuous Parameter Poisson Process"	205-219
KUHN, H. W., "Variants of the Hungarian Method for Assignment Problems"	253-258
LADERMAN, Jack, BECKMANN, M. J., "A Bound on the Use of Inefficient Indivisible Units"	245-252
LEWIS, R., NEELAND, F., GOURARY, M., "An Inventory Control Bibliography"	295-303
MARKOWITZ, Harry, "The Optimization of a Quadratic Function Subject to Linear Constraints"	111-133
MARLOW, W. H., ISBELL, J. R., "Attrition Games"	71- 94
NEELAND, F., LEWIS, R., GOURARY, M., "An Inventory Control Bibliography"	295-303
OSHIRO, Sieki, FENNELL, Joseph, "The Dynamics of Overhaul and Replenishment Systems for Large Equipments"	19- 43
REED, Ronald L., HALE, Jack K., "A Formulation of the Decision Problem for a Class of Systems"	259-277
ROLLOF, CDR Yngve (Royal Swedish Navy), "System Analysis and/or Common Sense"	11- 18
SAVAGE, I. Richard, "Cycling"	163-175
SMITH, Jack W., "A Plan to Allocate and Procure Electronic Sets Using Linear Programming Techniques and Analytical Methods of Assigning Values to Qualitative Factors"	151-162
SMITH, Wayne E., "Various Optimizers for Single-Stage Production"	59- 66

	Pages
THOMPSON, Gerald L., "Decision Making and New Mathematics"	141-149
VERHULST, Michel, "The Concept of a 'Mission'"	45- 57
WEISS, George H., "On the Theory of Replacement of Machinery with a Random Failure Time"	279-293
WOLFE, Philip, FRANK, Marguerite, "An Algorithm for Quadratic Programming".	95-110
WOLFOWITZ, J., KIEFER, J., "Sequential Tests of Hypothesis About the Mean Occurrence Time of a Continuous Parameter Poisson Process"	205-219
ZABEL, Edward, "Measures of Industry Capacity"	229-244

* * *

U. S. GOVERNMENT PRINTING OFFICE : 1957 O - 421661